# 道德部落

## 情感、理智和冲突背后的心理学

［美］乔舒亚·格林◎著
（Joshua D. Greene）

论璐璐◎译



# Moral Tribes

Emotion, Reason,
and the Gap
Between Us and Them

**哈佛大学心理学教授开创性著作**

被誉为亚里士多德之后，人类伦理学研究的新突破！

# 版权信息

书名:道德部落

作者:[美]乔舒亚·格林

译者:论璐璐

ISBN:9787508665177

中信出版集团制作发行

版权所有•侵权必究

# 前言 常识道德悲剧

在黑暗茂密的森林东边，有一个牧羊人部落，他们在公共草地上以放牧绵羊为生。这里的规则很简单：每个家庭拥有同样多的绵羊。管理公共用地的是长老委员会，由每个家庭派出的代表组成。在过去的几年里，委员会做出了很多艰难的决定。比如，有一户家庭选择喂养体型特别巨大的绵羊，以便占用更多的公共用地。经过几番激烈的争论，委员会终于阻止了这种做法。另外有一户家庭向邻居家的绵羊投毒，因而受到了严厉的惩罚。有些人认为处罚过于严厉，但也有人认为处罚力度还不够。不管怎样，这个东方部落经历种种磨难兴旺起来，有些家庭的繁盛胜过了其他的家庭。

在森林西边，生活着另外一个部落。那里的牧羊人也有一块公共草地。在那里，每个家庭可以拥有的绵羊数取决于家庭中的人数。他们也有一个长老委员会，这个委员会也做出了很多艰难的决定。有一户家庭人丁特别兴旺，家中有12个孩子，远远超过其他家庭。有些人抱怨他们占用了太多的公共资源。另外一户家庭拥有6个孩子，因不幸染疾，一年之内失去了5个孩子。有人认为，将他们的财产一下子削减一多半是不公平的，对这个家庭来说是雪上加霜。但不管怎样，这个西方部落也经历了种种磨难兴旺起来，有些家庭的繁盛胜过了其他的家庭。

在森林北边，也有一个部落。这里没有公共草地，每个家庭都拥有自己的草地，并用篱笆围住。每个家庭的土地大小不同，肥沃程度也不同。有些北方牧民比同伴更加聪明勤奋，他们用自己的积蓄从相对贫困的邻居那里购买土地，扩大自己的地盘。但也有一部分北方牧民运气不好，他们努力工作，但依然无法战胜病魔，失去了儿女和部分牲畜，日子也相对清贫。还有一些得到命运垂青的牧民，无需智慧和勤奋，从先人那里直接继承了大片的肥沃土地。在这个北方的部落里，长老委员会没有太多事做，只要保证牧民遵守承诺，不侵犯别人的财产就好。不同家庭之间巨大的贫富差距导致了频繁的争斗。每年冬天，都有一些北方人饥寒交迫而死。但不管怎样，这个北方部落经历种种磨难兴旺起来，有些家庭的繁盛胜过了其他的家庭。

在森林南边，还有另外一个部落。这里的牧民不仅共享草地，他们的牲畜也是共有的。这里的长老委员会非常忙碌，他们要管理部落的羊群，给人们分配工作，并进行监督管理。整个部落的劳动果实平均分配给每一位成员。由于一些部落成员比其他人更加聪明勤奋，部落里时常也会发生冲突。虽然长老委员会听到了很多有人消极怠工的抱怨，但总体来说，大多

数人都是勤奋的。有些人受到集体精神的感染，积极工作，有些人则是由于不愿被邻居指责，因此才努力工作。不管怎样，这个南方部落经历了种种磨难。总体来看，南方部落的家庭条件不如北方部落，但日子也还说得过去。冬天也不会有任何南方人因饥寒交迫而死去。

一年夏天，一场大火席卷了整个森林，把这里化为灰烬。接着又下起了倾盆大雨，没过多久，曾经的茂密森林就变成了广阔平缓、绿草茵茵的小山丘，成为理想的牧场。周围的部落都想占有这块草地，因而发生了很多冲突。南方的部落认为新草场是大家共有的，应当由大家共同使用。他们为此成立了一个新的委员会，对新草地进行管理，并邀请其他部落派出代表。北方的部落对这个建议嗤之以鼻。在南方人忙着制订计划的时候，北方部落的家庭已经在新草地上建起了房屋和石墙，把自己的羊群赶上了草场。东方和西方部落的很多家庭也纷纷在新草地上建屋放牧，只不过他们不像北方人那样急切，也有一些家庭派出代表，加入了新的委员会。

4个部落为了新草地争斗不休，很多人和牲畜为此丧生。小摩擦逐渐升级为流血冲突，又逐渐演变成殊死搏斗的战争。起初，一个南方人的一只羊跑到了北方人的地盘，北方人把羊还了回去。后来，另一个南方人的羊也跑到了这个北方人的地盘，这个北方人要求南方人付钱赎回这只羊。南方人拒绝付钱，北方人就把羊杀掉了。为了报复，南方人捉了北方人的3只羊并把它们杀掉了。然后北方人又杀了南方人的10只羊。于是南方人烧毁了北方人的农舍，还杀死了一个孩子。接着，10个北方家庭冲进南方人的集会地，放了一把火，杀死了很多南方人，其中还有不少孩子。就这样，双方冤冤相报，打斗不断，绿色的山丘都被血染红了。

更糟糕的是，还有很多远道而来的部落想在新草地上定居。一个部落声称新草地是神明对他们的馈赠。他们的圣书上早就预言过那场烧毁森林的大火和后来恢复生机的草地。另一个部落则说这块草地是他们祖先的家园。很多年以前，他们的部落被赶走了，但从这里还没有成为森林的时候起，他们的祖先就居住在这里。很多部落的规则和习俗在外人看来都十分怪异，甚至荒谬。比如：黑绵羊和白绵羊不能圈养在同一个围栏里；女人不能在公共场合裸露耳垂；周三不得唱歌等。曾有埋怨邻家的女人在照看羊群时裸露耳垂，而当时他的儿子也在场。但邻家女人拒绝遮盖自己的耳垂，这位尽责的父亲对此十分恼火。还有一次，一个女孩儿对一个男孩儿说，他们家信仰的神灵根本不存在。男孩儿在震惊之余将这件事告诉了自己的父亲。于是男孩儿的父亲便向女孩儿的父亲告状，但女孩儿的父亲竭力维护自己的女儿，表扬了女儿的聪明智慧，拒绝向男孩儿一家道歉。最终根据男孩儿所属部落的法律，这个女孩的父亲被杀死了。一场无休止的争斗又开始了。

尽管冲突时现，但入驻新草地的牧民们有着许多的共通之处。大部分人的

愿望都是一致的：健康的家人，美味营养的食物，舒适的房屋，省力的工具，与朋友家人共度的闲暇时光。所有的牧民都喜欢音乐，也喜欢听英雄和坏人的故事。即使是相互争斗的双方，他们的心理状态也大同小异。一旦遇到他们认为不公正的事，愤怒和厌恶的情感便会产生。各方争斗的动力都是为了维护自身利益、维护公正。牧民们不仅会为自己而战，也会为家庭、朋友和同族人而战。他们在战争中心怀荣誉感，反之则会为自己感到羞愧。他们坚决维护自己的声誉，基于行动对他人做出判断，喜欢相互交流看法。

新草地上的各个部落虽然各有不同，但也有一些核心价值能够得到普遍认同。没有哪个部落能够容忍绝对的自私，也没有哪个部落要求他们的成员完全无私。即使是在共享牲畜的南方部落，牧民们结束一天的工作后，也可以为自己谋些福利。没有哪个部落允许普通人撒谎、偷盗或任意伤害其他人。（也有一些部落中存在特权阶级，可以任性而为。）

入驻新草地的各个部落彼此争斗，无休无止，经常有人伤亡。但在各自的观念中，他们都是道德的人。这些争斗不代表他们本性自私，而是因为他们对于道德社会的看法不同，彼此不能相容。尽管来自不同部落的学者之间会产生争议，但争斗的起源并非学术争论。事实上，每个部落特有的哲学观点已经渗入了日常生活，每个部落对于道德常识也都有自己的界定方式。新草地上的住民们并非不道德，争斗之所以会发生，是因为不同部落对于新草地上应有的生活状态的道德视角不同。我把这种情况称为“常识道德悲剧”。

新草地的寓言当然是虚构的，但常识道德悲剧却是现实存在的。现代生活中，造成各种分歧的道德问题背后，便是这个深层次的核心悲剧。本书阐释了我们应当如何理解并最终解决这些问题。与其他畅销书的作者不同，我不保证你能够通过阅读本书解决个人问题。我希望向大家呈现的是明晰，有了明晰的认知，我们才有动力和机会找到同类，进行合作。

本书对道德进行了完整的阐述，包括何为道德、道德是如何产生的，以及道德是如何深入人心的。书中解释了道德问题的基本构成，也解释了大脑生来能够解决的问题和我们现在面对的社会问题有何不同。最后，我希望本书能够使大家从全新的角度认识道德，能够将这种伦理学进行推广，获得所有人类部落的认可。

这本书的目标是远大的。我从20岁左右便开始构思，这些想法引领我接触了哲学和科学这两个相互交织的学科。我从古代先哲身上获得了写作这本书的灵感。在道德认知这个新领域，我利用实验心理学和认知神经科学的知识对道德思维的结构进行阐释，本书就是基于这个领域的一些研究而著成的。此外，我还借鉴了上百位科学家的研究成果，关于人类如何做决

策，文化背景和生物学因素如何影响人们的选择等方面，他们都有杰出的发现。在这本书中，我试着将这些知识进行整合，将这门自我认知的科学转变为一门实用的哲学，用来指导我们解决更大的问题。

## 新草地上的生活

在巴拉克·奥巴马总统的第一届任期内，最大的两个问题是：医疗和经济。这两个问题都反映出了北方牧民的个人主义和南方牧民的集体主义之间的矛盾。“患者保护与平价医疗法案”（又称“奥巴马医改”）旨在为美国建立国家医疗保险体系。自由派举双手赞成，倒不是因为这个计划完美无缺，而是因为这个体系是美国向正确方向迈出的历史性一步。美国终于同现代的其他国家一样，开始向其民众提供基本医疗保险。然而，人数众多的保守派则极度鄙视奥巴马医改计划，他们认为这种改革是走向社会主义和自我毁灭的第一步。最近关于医改的争论中充斥着谬误[1]，但在谎言和半真半假的言语中，哲学上的分歧的确是真实存在的。

这个分歧的核心和其他很多分歧十分类似，都是个人权利和（所谓的或真正的）更多人的利益之间的冲突。美国医疗保险要求每个人出资参保，可能是个人自掏腰包，也可能是通过税收购买医保。保守派从法律的角度对奥巴马医改计划提出质疑，最终不得不由最高法院对医改计划进行历史性裁决。最高法院肯定了奥巴马医改计划的合法性，因为这个计划通过自愿购买和税收（符合宪法规定）进行筹资，并非政府强迫人民购买（可以说是不符合宪法规定的）。但从法律角度来看，税收筹资和强制购买之间的界限不过是一个技术问题。反对奥巴马医改计划的人并不在乎它的筹资方式是强制购买还是强制税收，他们在意的是“强制”这件事。或许奥巴马医改计划不能算是社会主义，但它所包含的集体主义元素确实超出了很多人的接受范围，因为这种计划打着更多人利益的旗号，限制了个人的自由。

2012年美国总统大选过程中，共和党内初选时，所有候选人都尽己所能不断指责奥巴马，将他称为“共产主义分子”，发誓要赶他下台。初选辩论时，记者沃夫·布雷泽采访了得克萨斯州众议员罗恩·保罗（ron Paul）。

> 布雷泽：假设有一位30岁的健康男子，有一份不错的工作，生活十分安逸。如果他现在决定：“我不愿每月花两三百美元来支付医疗保险。因为我十分健康，根本不需要医保。”但不幸的是，如果他突然昏迷不醒，迫切需要医疗保险，那么谁来支付这笔费用呢？
>
> 保罗：当今社会中，人们能够接受福利主义和社会主义，自然会希望由政府来照顾这个人。

布雷泽：那么，你希望由谁来负责呢？

保罗：这位男子应当做自己想做的任何事，为自己负责任。我给他的建议是，主动去买一份高额医疗保险，而不是被迫去买……

布雷泽：但他没有买保险。他没有保险，却需要在重症监护病房待半年。谁该为其埋单呢？

保罗：为自己负责，这就是自由的真谛。如果你想为每个人做好万全的准备……（掌声）

布雷泽：但是议员先生，您的意思是不是说社会应该让这个人自生自灭呢？

就在保罗迟疑的时候，人群齐声呼喊："对！自生自灭！"这就是北方牧民。保罗无法对这种观点表示赞同，也无法表示反对，只好回答说，这个人的邻居、朋友和附近的教堂应该照顾他。虽然保罗没有明说，但他暗示道，如果没人愿意且有能力为这个人支付医疗费用，那么政府就应任其自生自灭。如你所料，偏南的牧民对此持反对态度。

（注：在新草地寓言中，南方牧民是极端的集体主义者和社会主义者，他们的思想比当代的主流自由派思想更加偏左（针对这一点的争论时有发生）。因此，在当代政治的语境下，我会将当代自由派称为"偏南的"而不是"南方的"。相比之下，当代美国的保守派与故事中的北方牧民则更加相像。）

奥巴马总统的第一届任期内，不景气的美国经济和医疗保险问题都是政府关注的重点。2009年奥巴马初上任时，金融业在房地产市场上投入巨额赌注，而房地产泡沫经过10年膨胀终于破裂，美国经济直线下跌。美国政府采取了一系列措施，试图阻止金融系统全面崩溃。首先，2008年年末布什总统在任时，联邦政府出资拯救了几家深陷危机的投行。*奥巴马总统上任后，政府帮助汽车行业摆脱困境，并向无力偿还房贷的人们伸出援手。北方牧民对这些措施提出了不同程度的反对意见，他们认为政府应当允许银行、汽车制造商和绝望的房主们"自生自灭"，就像罗恩·保罗假设中的病人一样。他们问道，美国的纳税人凭什么要为这些人所犯下的错误埋单？在救助不负责任的决策者这件事上，偏南的牧民并不十分热衷。但他们认为，为了更多人的利益，为了不让整个经济因为一些人的错误选择而崩溃，这些措施是必需的。奥巴马上任一年时间里，民主党在国会中通过了他提出的7.87亿美元经济刺激计划，即《2009年美国复苏及再投资法案》。当然，一贯支持减少政府开支、削减税收的北方牧民对此法案表示

反对。他们认为，与其这样，还不如把这些钱分给个人，由个人自行支配。

与医疗和经济问题相关的更大问题是贫富差距。2011年的“占领华尔街”抗议活动将这个问题推到了人们面前。1979~2007年，美国最富有家庭的收入猛增。最富有的1%的人群收入增长达到275%，而美国大众的收入增长仅为40%。（金字塔尖的人群，即前0.1%的富人，收入增长率更高，约为400%。）“占领华尔街”活动的口号“我们是那99%”就来源于此，人们呼吁美国进行经济改革，重建一个更加平等的财富和权利分配体系。

面对不断拉大的收入差距，人们持有两类观点。崇尚个人主义的北方牧民认为，胜者为王，公平合理，失败者无权抱怨。华尔街的反示威标语写道：“占领工作台！”总统竞选者赫尔曼·凯恩（Herman Cain）把示威者称为“非美国人”，而共和党最终提名的候选人米特·罗姆尼（Mitt romney）则指责示威者发动了“阶级战争”。

2012年9月，自由派杂志《琼斯夫人》（*Mother Jones*）爆出猛料，堪称美国竞选史上最重磅炸弹。该杂志将罗姆尼的一段秘密谈话录音通过网络进行了披露。谈话录音中，罗姆尼称美国近一半的人口永远“不愿承担个人责任，不愿关心个人生活”，故意依赖政府。这段流传甚广的录音中，罗姆尼还认为，这“47%”的人口收入太低，除了工资税之外根本达不到缴纳个人所得税的标准，他们只配过现有的生活。

偏南的牧民对收入差距持有不同态度。他们认为，富人们操纵国家系统，制定对自己有利的政策。例如，投资收益享受的低税率、种种税收漏洞，以及海外避税港都使得像米特·罗姆尼这样的富人得以享受低于多数中产阶级的税率。此外，最高法院对“公民联合会诉联邦选举委员会案”进行裁决后，对于“独立”政治团体的竞选资助不再有上限规定，富人们获得了前所未有的自由，可以随心所欲地收买选票。偏南的牧民还认为，即使不公正的操纵不复存在，要想维持社会公正，财富的流动和重新分配也必不可少。否则富人将会利用财富优势变得更加富有，并将财富优势传递给下一代，其子女一出生便占尽优势。如果不对财富进行重新分配，社会将分化为两极，形成固定的富人阶级和穷人阶级。

马萨诸塞州议员伊丽莎白·沃伦第一次参与选举时，曾在一次演说中提出了关于财富分配的一个“偏南”观点，并在视频网站YouTube上迅速走红：

> 在这个国家，所有富人的发展都不是单凭一人之力完成的。无一例外。假设你开了一家工厂，这很棒。但我需要提醒你：当你将货物运向市场时，你使用的公路是我们所有人出钱修建的；你雇用的工人是

> 靠我们所有人出钱完成学业的；你在工厂里十分安全，这是因为我们所有人出钱设立了公共安全和消防部门，你不必担心会有抢劫团伙把工厂洗劫一空……看，你开了一家工厂，可它变成了一个庞然大物——愿主保佑！当然，工厂的绝大部分都属于你。但默认的社会规则是，你可以占有绝大部分，但作为回报，你需要出资帮助后来者。

罗恩·保罗回击了沃伦，将她称为社会主义分子，并说由她执掌的政府什么都做不了，只能“拿着枪对人民行窃、抢劫，将一个人的财富强行分配给另一个人”。保守派评论员拉什·林博的态度则更加强烈，他将沃伦称为共产主义分子，说她是“一只憎恨宿主的寄生虫”。

不同部落之间还有很多其他方面的分歧，但这些分歧与个人主义和集体主义之间的根本分歧并没有明显的关系。在美国，“是否应当采取措施应对全球变暖？何种措施更为合适？”这样的问题引发了激烈的争论。这种争论的本质看似与价值观无关，争论的焦点不过是一些事实：“全球变暖的威胁是否真实存在”以及“全球变暖是否应当归咎于人类”。但这些争论的产生真的只是由于人们解读数据的方法不同吗？相信全球变暖的人认为，所有人都必须做出牺牲（少使用燃料、缴纳碳排放税等等），以保障人类的共同利益。个人主义者出于本能质疑这些要求，而集体主义者则更倾向于相信这种观点。这说明，价值观很可能影响到我们对于事实的认知。

我们在新草地上遇到的麻烦并不是个人主义和集体主义本身的对立，问题的产生与人类思维中集体主义思想的界定有关。从某种程度上说，几乎所有人的思维中都有集体主义思想存在。只有隐士能够算作纯粹的个人主义者。回想一下罗恩·保罗对那个没有医保的男子的态度。保罗并没有说我们应当任由这位男子死去，他说的是这位男子的朋友、邻居和附近的教堂应该照顾他。这就说明，不同之间的分歧并不一定是个人主义和集体主义的对立。差别在于内部的集体意识是强还是弱；部落居民更愿意从“我们”还是“他们”的角度看世界；对美国联邦政府和联合国等部落间集体机构持有怎样的态度。对于保守派而言，他们不过是将“我们”的范畴界定得更小而已。

有些部落价值观的本质是从局部着眼，甚至只关注该部落本身，因此会与其他部落产生分歧。有些部落给某些特定的神灵、领袖、文本或行为（可将其称为“专有名词”*）赋予了特殊的权威。例如，很多穆斯林相信，不论是否信奉伊斯兰教，任何人都不能创造先知穆罕默德的视觉形象。有些犹太人相信犹太民族是上帝“拣选出的子民”，他们遵照神圣的旨意拥有以色列。很多美国的基督徒认为公共场合的大楼上都应张贴《摩西十诫》，所有美国人都应宣誓效忠“上帝治理下的国度”。（这里所指的可不是印度的毗湿奴神。）

有些部落的道德行为十分随意，或是看似随意，但这些部落大多不会将自己的道德规则强加于人，至少在发达国家是这样的。比如，正统犹太教徒不会要求非犹太教徒不吃龙虾，也不会要求他们给男孩子行割礼。天主教徒不会要求非天主教徒一定要在大斋首日用灰在额头上画十字架。最终引发公开辩论的部落分歧大多是与性相关的话题（如同性婚姻、军队里的同性恋、政府官员的性生活等），或是界限模糊的死亡问题（如堕胎、安乐死、科研活动中使用胚胎干细胞等）。这些问题之所以会成为道德问题，绝非偶然。我们可以把性和死亡看作控制部落人口增长的油门和刹车（比如，同性性行为和堕胎都是繁衍后代过程中人们做出的不同选择）。但我们不明白的是，为什么不同部落对于性、生命和死亡的态度迥然相异，为什么有些部落更愿意将其观点强加于外人。

在我写书的同时，美国在新草地上的旅程也在急速推进。如果你在稍晚的时代或是在不同的地区读到这本书，你所面临的具体问题可能会与书中不同，但问题背后的紧张关系却是大同小异。环顾周围，你会发现北方牧民和南方牧民正在为政府权限的大小争论不休；不同部落为了“我们”这个概念的范围各执己见，在性和死亡等道德问题上苦苦争辩，为维护各自部落专有名词的神圣不遗余力。

## 世界伦理学的讨论

如果你是一位来自其他星球的生物学家，大约每一万年拜访地球一次，来观察这个星球上生命的进化历程，在你的观察记录本上大概会留下这样的文字：

现代人：有大脑，直立灵长类，声音语言，时有进攻性

| 观察次数 | 人口数目 | 记　录 |
| --- | --- | --- |
| 1 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 2 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 3 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 4 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 5 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 6 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 7 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 8 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 9 | <1 000 万 | 以狩猎采集为生的族群，使用原始工具 |
| 10 | >70 亿 | 全球工业经济，高科技武器/核武器，电子通信，人工网络，太空旅行，大规模社会/政治机构，民主治理，先进的科学探索手段，读写能力普及，艺术水平高超（见附录） |

从人类开始出现直到一万年前，人类似乎并没有取得很大的进步。然而现在，我们坐在温度可控、使用人工照明的房间里，阅读和书写着关于我们自己的书籍。我们所取得的进步不仅仅是单纯的物质享受。很多人哀叹人类文明的退步，但事实却恰好相反，人类在逐步发展，人与人之间的相处也日渐融洽。在人类发展史上，暴力行为逐渐减少，即使把近代历史纳入考量范围也是如此。现代市场经济并没有将我们变为斤斤计较的吝啬鬼，相反，人性的善良在这一过程中得到了扩展。

但是，我们也有很多不足之处。20世纪是历史上迄今为止最为和平的年代（考虑到人口增长的幅度），但在这期间，依然有大约2.3亿人口在战争和各类政治冲突中丧生，这些人的尸体排起来足以绕地球7圈。新世纪中，这一数字仍在增加，尽管增速有所减缓。例如，当下在达尔富尔发生的冲突已经造成30万人死亡，有些人死于直接冲突，有些则死于各类疾病。还有约占地球总人数1/7的10亿人生活极端贫困，他们拥有的资源少得可怜，不得不为生存而不停奋斗。还有超过两千万人所从事的工作并非出于自愿（例如：奴隶），他们中多数都是被迫卖淫的妇女和儿童。

即使是在世界上相对幸福的地区，生活对于大部分人来说也绝非公平。美国的研究员曾给不同的招聘人员发送内容完全相同的简历，唯一的区别就是有些简历上的人名看上去像是白人的名字（比如艾米莉、格雷格），而有些简历上的人名看上去像是黑人的名字（比如拉齐莎、贾马尔）。结果写有白人名字的简历接到的招聘电话比写有黑人名字的简历多了50%。最糟糕的是，人类通向和平繁荣的努力面临威胁：环境恶化和大规模杀伤性武器蔓延。这两大问题可能会对这一进程造成严重干扰，甚至会造成历史

的倒退。

面对如此黯淡的前景，这本书的写作却基于一个相当乐观的前提：通过改进人类看待道德问题的方式，我们有能力创造出更加和平繁荣的未来。在过去的几个世纪里，新的道德观念占据了人们的思想。多数人相信，任何人类部落都不应比其他部落享受更多特权；人人都有权享受基本权利和自由；暴力是解决问题的下下策。（也就是说，有些部落的部族意识已然大大降低。）尽管我们对这些观点的赞同更多停留在理论层面，而没有在实践中表现出来，但仅仅是表示赞同的态度，就已经是人类历史上前所未有的事了。历史告诉我们，人类已经取得了很大的进步，不仅是技术进步，更包含道德进步。

史蒂芬·平克（steven Pinker）从另一个角度就当下的道德问题提出疑问：我们正在做的什么事是正确的？如何才能做得更好？我想，我们需要的是一个全球统一的伦理学体系，一个能够化解不同部落间道德冲突的体系。这个想法并不新鲜，从启蒙运动时起，历代道德思想家都梦想着建立这样一个体系，可惜从未实现。我们拥有的只是一些共同的价值观、一些不同的价值观和一些共同的法则，还有一个共同的词汇表，用于描述我们彼此赞同或反对的价值观。

对道德的理解包含两个前提：第一，我们必须理解当代道德问题的构成，理解这些问题和人类大脑生来能够解决的问题有何不同。在本书的第一部分，我们会对此进行阐释。第二，我们必须理解人类道德思维的结构，了解不同的思维模式分别适于解决何种问题。这是第二部分的主要内容。在第三部分中，我们对道德问题和道德思维的理解将被付诸实践，寻找一种解决方案，初步建立全球伦理学体系。在第四部分中，我们将对一些有力的反对意见做出回应，在第五部分中，我们将这个理论体系应用到现实世界中。下面我将对本书结构做进一步的详细介绍。

## 本书结构

第一部分（“道德问题”）对两类主要的道德问题进行辨析。第一类问题较为基础，是“我”与“我们”之间的对立，即利己主义与关心他人之间的对立。我们的道德思维生来就能解决这类问题。第二类道德问题具有鲜明的时代色彩，是“我们”与“他们”之间的对立，即我们的利益和价值观与他人的利益和价值观之间的对立。这个问题在本书开篇第一个隐喻中便已经得到阐释，“新草地”的寓言所反映的就是常识道德悲剧。（诚然，“我们”与“他们”之间的对立由来已久。但这种对立在历史上一直都属于策略问题，而非道德问题。）相较于不同部落间的道德纷争，这个问题则更加严重。第一部分中，我们将探索人类大脑中的道德机制如何解决第一类问题

（第2章），第二类问题又是如何产生的（第3章）。

在第二部分（“道德反应的快与慢”）中，我们将深入探索道德思维，引入本书的第二个隐喻：道德思维像是一个可以在两种模式间切换的照相机，一种是“傻瓜型”的自动模式（比如“人像”或“景物”模式），另一种是手动调整模式。自动模式非常高效，但相对死板。手动模式十分灵活，但相对低效。道德思维的自动模式就是第一部分中提到的道德情感。这种情感更像是一种本能，让我们在个人关系和小团体中维持合作。相比之下，手动模式则好像一种实践推理能力，能够解决道德问题以及其他的实际问题。在这一部分中，我们将讨论情感和理智如何塑造道德思维（第4章），以及道德的这种“双加工模式”如何反映人类思维的整体结构（第5章）。

在第三部分中，我们将引入本书的第三个，也是最后一个隐喻：通用货币。我们将开始对元道德的探寻。元道德是全球统一的伦理学，能够在不同部落的道德观发生冲突时进行裁决，就像一个部落的道德观可以裁决部落内的个人利益冲突一样。元道德可以对相互冲突的部落价值观进行衡量取舍，这一过程中便需要有统一的体系，即通用货币，来衡量价值。在第6章中，我们会向大家介绍一种元道德，也是常识道德悲剧的一种解决方案。在第7章中，我们会讨论建立通用货币的其他方式，最终发现我们并没有太多的选择。在第8章中，我们会对第6章介绍的元道德做进一步分析。第6章提出了一种被称为“功利主义”的思想。本章则介绍了功利主义如何诞生于易于理解的价值观和推理过程，又如何为我们提供通用货币。*

多年来，哲学家们凭借直觉提出了有力的观点，反对功利主义。在第四部分（“道德信念”）中，我们将基于新产生的道德认知，重新考量这些观点。一旦我们对道德的“双加工模式”有了更深的理解，功利主义观点将会变得更加有趣（第9章和第10章）。

最后，在第五部分（“道德出路”）中，我们将回到新草地，回到我们最初想要解决的现实问题。既然前文已经为功利主义正名，这一部分便会将其付诸实践，并为它取个好名字。事实上，“深度实用主义”（第11章）比“功利主义”更加恰当。从积极的角度，也是我们更加熟悉的角度来看，功利主义是务实的，它提倡灵活、实际，乐于相互妥协。但同时，功利主义并不是权宜之计，而是一门深刻的哲学，是通过寻找共同的价值观——通用货币——来解决分歧的一种方式。

我们将讨论：成为一名实践中的深度实用主义者意味着什么。我们何时应当信任道德直觉，使用自动模式，又应当在何时切换到手动模式？切换到手动模式时，我们应当如何利用推理的力量？我们可以利用大脑为感性的道德信念寻找合理化解释，也可以选择突破部族主义本能的局限。我本人

选择突破，突破“傻瓜型”的道德本能，选择换一种方式思考并谈论我们之间的分歧。在第12章中，我将为新草地上的生活提出6条简单实用的规则，以此作为本书的结尾。

---

[1]这是本书中唯一的脚注。尾注和直接的引文中有很多补充信息。当下很多书籍并未在书中标明尾注对应的文本。我也不想让上百个编号影响读者的观感，但我还是希望读者能够了解书中何处会出现有趣的信息，补充的信息量大致有多少。因此我借鉴了亚洲菜单上标注辣椒辣度的方式，在本书中使用如下方法进行注释：用星号代表注释内容的长短度（*代表较短，约几句话；**代表适中，约几段话；***代表较长，约几页内容）。只写出引用出处的注释不在文中另行标出。（想要了解更多关于“充斥着谬误”的内容，请参看文后注释中的介绍。）

# 第一部分 道德问题

# 第1章 公地悲剧

也许你已经注意到，新草地的寓言是一个系列。这个寓言最初由生态学家加勒特·哈丁（Garret Hardin）提出。他在1968年发表了一篇著名的论文，题为《公地悲剧》。在哈丁的寓言中，一群牧羊人共同享有一块公共草地。这块公共草地面积很大，足以养活很多牲畜，但牲畜数量不能无限增多。牧羊人经常会面临是否增加畜群数量的选择。理性的牧羊人会怎样做呢？如果在现有畜群的基础上增加一只羊，牧羊人以后在市场上就可以多卖一只羊，获得的收益也会增加很多。但养活这只羊的成本则需由使用公共草地的所有人共同承担。这样一来，牧羊人增加一只羊所需的成本很低，获得的收益却很高。因此，在不超出公共草地承载能力的范围内，畜群的数量越多，对牧羊人就越有利。当然，每位牧羊人的情况都是一样的。如果每位牧羊人都基于个人利益做决定，这片公共草地就会被彻底毁掉，所有人都会一无所有。

哈丁的“公地悲剧”阐释了合作的问题。合作并不总是困难的。有时候，合作是水到渠成的选择，有时候却难于登天。介于两者之间的情况，则是“有趣的合作”。

假设有两个人，一个叫亚特，一个叫巴德。两人同处于海上的一艘划艇中，在巨大的风浪中努力保持平稳。除非两人都拼命划船，否则谁都无法幸存。在这种情况下，个人利益和集体利益（这时，“集体”指的就是这两个人）完全一致。对亚特和巴德来说，对“我”最好的选择也就是对“我们”最好的选择。然而在另一种情况下，合作则变成天方夜谭。假设亚特和巴德的划艇正在下沉，但两人只有一件救生衣，且救生衣无法共享。这时，“我们”的概念不复存在，只有两个单独的“我”。

在以上两个例子中，合作或是必然选择，或是天方夜谭，并不涉及社会问题。只有像哈丁的寓言那样，个人利益和集体利益既不完全重合，也不完全冲突时，合作才成为一个问题：有一定难度，但也不是完全无解。让我们再次回到哈丁的寓言，每一位牧羊人都能够通过增加畜群数量而获得更多利益，但这种做法会导致集体利益受损，对任何人来说都不是最好的选择。因此，合作就是在可能的情况下，将集体利益置于个人利益之上。合作是社会存在的核心问题。

为什么一个人要存在于社会之中？为什么不能索性自己生活？答案是，有时候群体可以完成一些单凭个人无法完成的事情。这条原则从地球上生命诞生开始，就一路引导着生命的进化。大约40亿年前，分子聚在一起，形成了细胞。大约20亿年前，细胞聚在一起，形成了更加复杂的细胞。接

着，10亿年之后，更加复杂的细胞聚在一起，形成了多细胞生物。这些群体之所以能够进化，是因为其中的各个组成部分都能够协同工作，将基因以新的、更加有效的方式进行传递。再将时间快进10亿年，来到我们所生活的世界。这个世界中生活着无数的群居动物，从蚂蚁到狼再到人类。同样的原则依然适用。蚁群和狼群能做的很多事，一只蚂蚁或一匹狼都无法做到。人类则通过相互合作，成为统治地球的生物。

人与人之间绝大多数合作都是“有趣的合作”，以个人利益和集体利益的部分重合为前提。在关于亚特和巴德的第一个例子中，我们假定两人的利益完全一致：两人都必须拼命划船，否则都会被淹死。但类似这样的情况并不多见。更加典型的情景是，亚特和巴德中有一个人可以不那么用力划船，划艇依然能够保持平稳。普遍来看，绝大多数的合作模式中，个人都能找到机会牺牲集体利益、扩大个人利益。也就是说，在合作过程中，个人利益和集体利益之间，也就是“我”和“我们”之间的冲突几乎不可避免，只是程度会略有不同。因此，几乎所有的合作都有被侵蚀的风险，正如哈丁寓言中的公共草地。

个人利益和集体利益之间的矛盾十分常见，甚至在很多看似不涉及合作问题的场景中，这对矛盾也会出现。假设亚特正在美国西部的荒原上漫游，独自走在一条荒凉的山间小路上。在远处的山上，亚特看到了另外一位旅行者的身影，他也是独自一人，正在翻越前方的山脊。这个人带武器了吗？亚特无法判断。但亚特自己是有武器的，他还是一位不错的射手。亚特举枪瞄准那位陌生人，自信能一枪毙命。但亚特应该开枪吗？从自私的想法来看，开枪对亚特没有任何损失。如果他将那位陌生人杀死，就不必担心会被抢劫了。所以，开枪射杀那位陌生人符合亚特的私利。

同时，在同一片区域跋涉的巴德也面临着同样的选择。他需要横越山脉，取回之前藏好的金子。在路边，巴德遇到了一位熟睡的陌生人。他知道自己返程时很可能会再次与这个人相遇，而那时，自己身上将会带着金子。这位陌生人会对他实施抢劫吗？巴德无法判断。但巴德知道，如果他在陌生人的威士忌里下毒，熟睡的陌生人是不会察觉的。

从个人利益出发，便会得到这样的结果：巴德在亚特的威士忌里投毒。几个小时后，亚特击毙巴德。再过几个小时，亚特喝下威士忌，也将死去。假如亚特和巴德都能够对陌生人的利益稍微多点考虑，两人就都能够活下来。然而，就像哈丁寓言中的牧羊人一样，他们都被个人利益的诱惑击败了。结论：即使是最基本的尊重和不侵略，也是一种合作。不论是人类还是其他物种，都不能将这种行为视作理所当然。黑猩猩是与人类最相似的两种动物之一。假设两群雄性猩猩在小径上相遇，其中一群的数量明显占

优势，那么这群黑猩猩很可能会将另一群猩猩杀掉。既然有这种能力，为什么不做呢？谁会需要竞争？由此看出，和平就是合作的难题之一。

几乎所有的经济活动都面临合作的问题。你从商店里买东西时，会相信店员递给你的就是你要买的东西（比如你买到的确实是碎牛肉，不是碎鼠肉）。同样的，店员相信你递给他的十美元钞票是真的（不是假钞），也相信你不会把没付钱的商品装进自己的口袋。社会中存在法律与警察，其存在的意义就是保证人们履行承诺。几乎所有的经济活动都会涉及“有趣的合作”，都会涉及个人利益和集体利益之间的竞争，我们需要额外的机制来维护秩序。

在市场之外，几乎所有的人类关系都需要有取有予。若一方或者双方索取太多、付出太少，关系往往会破裂。事实上，个人利益和集体利益之间的矛盾不仅出现在人与人之间，也出现在人的身体内部。如前所述，复杂的细胞已经相互合作了10亿年。然而，动物体内也有些细胞会为自己谋取福利，而不是为整个团队着想，这种现象并不稀有，我们将其称为癌症。

## 道德的功能

达尔文逝世后，人类道德成为科学谜题。自然选择理论可以解释灵长类动物为何会进化成智商高、直立行走、有语言、毛发不多、有两足的样子。但人类的道德究竟来自何处？达尔文自己对这个问题也百思不得其解。人们曾经认为，自然选择鼓励冷酷的自利行为。能够霸占所有资源、击倒对手的个体才能活得更好，繁衍更多的后代，将冷酷自私的血统传遍世界。若果真如此，道德是如何在这个被丁尼生（Tennyson）称为“爪牙染血”的世界中发展的呢？

现在我们终于有了答案：道德的发展是为了解决合作的难题，是为了避免“公地悲剧”。

道德是人类不断对心理做出适应性调整的产物，它能够使性本自私的个体享受到合作的果实。

道德如何实现这一目标？下一章将会做出详细的解释。概括而言，道德的本质是利他、无私的，是一种甘愿牺牲个人利益成全他人的思想。自私的牧民会不停增加畜群数量，直到自己再也无法获得更多利益。我们已经知道，这种行为最终只会走向毁灭。然而，道德的牧民明知自己会承受损失，也可能会出于对他人利益的考虑，自愿限制畜群的数量。因此，对一群道德的牧民来说，只要他们愿意将“我们”的利益置于“我”的利益之上，

就能够避免公地悲剧，共同繁荣发展。

道德为促进合作而发展，但这个结论基于一个重要的前提。从生物学角度看，人类的进化就是为了合作，但这种合作仅限于特定的人群。人类道德思维的发展能够促进群体内部的合作（或者说是人际关系网内部的合作），但这种发展并不利于群体之间的合作（至少不是所有群体之间都能够相互合作）。这种机制是如何被发现的？为什么道德的发展不能在更高层面上促进合作？因为普遍合作与自然选择下的进化机制相悖。我希望事实并非如此，但面对现实，我们无处回避。（此处需要补充一点，这个结论并不意味着人类注定无法开展普遍合作。稍后会就这一点深入阐释。）

进化的本质是竞争的过程：狮子跑得越快，捉到的猎物就越多，繁衍的后代也越多。因此在下一代中，跑得快的狮子所占比例也会上升。如果资源不必通过竞争获得，这种情况便不会出现。如果狮子的食物取之不尽，跑得快的狮子便失去了优势，下一代狮子的平均奔跑速度也不会比上一代有所提高。根据自然选择原理，没有竞争就没有进化。

同样的道理，如果合作的倾向不能使合作双方在竞争中占据优势，那么这种倾向便无法进化（生物学角度）。假设有两群牧民，一群之间相互合作，另一群则相反。相互合作的牧民将各自的畜群限制在一定规模，保护了公共草地。牧民们长期的食物来源也有了保障。相互不合作的牧民以个人利益为行为准则，不断扩大各自的畜群。最终，公共草地被慢慢毁掉，牧民们的食物也所剩无几。因此，第一群牧民因合作而受益，可以将第二群牧民取而代之。他们可以坐视第二群牧民因饥饿而死。如果他们略有野心，也可以发动一场没有悬念的战争——吃饱喝足与腹中空空的对阵。一旦相互合作的牧民占据主动，他们便可以豢养更多的牲畜，养活更多的孩子。因此下一代牧民中，相互合作的牧民比例也会增加。合作之所以能够进化，不是因为它有多么“美好”，而是因为合作能够提供优势，有利生存。

在肉食动物的进化过程中，竞争对合作的发展至关重要。假设两群牧民生活在一片神奇的草地上，这片草地能够养活无限多的动物。这种情况下，不愿合作的牧民毫无劣势可言。自私的牧民可以随意扩大自己的畜群，结果不过是畜群数量不断增加。合作得以发展的前提是，有合作倾向的个体能够在竞争中胜过没有合作倾向（或不愿合作）的个体。因此，如果说道德是人类为适应合作而做出的选择，那么我们今天之所以拥有道德感，只是因为我们的祖先拥有道德感，在竞争中胜过了道德感不及他们的邻居。因此，在生物学适应性理论的范围内，道德的发展要求人们在“我们”和“我”之间优先考虑“我们”，在“我们”和“他们”之间也优先考虑“我们”。（注意，我并没有假定道德通过群体选择进化而来。*）这种思维方式对人类发展影响深远。

“道德作为群体之间的竞争策略发展而来”这个观点似乎有些难以接受，原因至少有二。首先，大部分的道德行为似乎与群体间的竞争拉不上关系。比如在堕胎问题上，维护选择权还是维护生命权与群体间的竞争有什么关系呢？类似的例子还包括人们对同性恋婚姻的态度、对死刑的态度，以及拒绝食用某些食物等。接下来的章节中，我们会谈到，道德思维与群体间竞争的联系是间接的、隐性的。这里先将这个问题暂时搁置一下。

第二，将道德视为击败“他们”的策略听上去有些是非不分，甚至有些不道德。为什么会这样呢？要知道，道德的实际作用已经超越了其本身进化（生物学角度）的目的。明晰了这一点，这种矛盾的感觉也就迎刃而解了。作为道德的存在，人类拥有的价值观可能会与最初促使道德产生的力量背道而驰。此处借用维特根斯坦的一个著名比喻：道德爬上了进化的梯子，然后把梯子一脚踢开。

我们可以用节育手段的发明进行类比。人类进化出庞大复杂的大脑，可以发明创造，进行技术革新，解决复杂问题。总体来说，人类解决问题的能力有利于繁衍并养活更多的后代。但在节育问题上，人类开动脑筋，将聪明才智用于限制后代数量，挫败了大自然的“意图”。*同样的，我们也可以将道德引向一个全新的方向，一个大自然从没“计划”过的方向。比如，我们可以为远方的陌生人捐款，而完全不期待任何回报。从生物学角度看，这种情况不过是小小的乌龙而已。但作为人类，作为有能力踢开进化之梯的道德存在，这种结果正是我们所想要的。道德是超越了进化目的而存在的。

## 元道德

人类面临两大道德悲剧的威胁。传统的“公地悲剧”源于自私，因为人们在“我们”和“我”的选择中没能优先考虑“我们”。道德是大自然解决这一问题的方法。现代的新悲剧叫作“常识道德悲剧”，也就是新草地上的居民遇到的问题。毫无疑问，我们需要道德来解决这个问题，但同时道德本身也是问题起因的一部分。在这个现代悲剧中，促进群体内部合作的道德思维恰恰是破坏群体间合作的元凶。在新草地的部落内部，牧民们因为相同的道德理想而紧密团结，然而部落之间却以不同的道德理想作为区分。这是十分不幸的情况，但基于前文得出的结论，这种情况却也无可避免，因为道德的发展本不是为了促进普遍合作。事实上，道德是作为群体间竞争的获胜策略而得以发展的。换句话说，道德发展的目的是为了避免“公地悲剧”，而不是避免“常识道德悲剧”。

那么我们作为现代牧民，能为此做些什么呢？这正是本书将要回答的问题。人类如何使自己的道德思维适应当今世界？是否存在一种道德思维，

能够使人类平静快乐地共同生活？

道德是大自然促进群体内部合作的方法，它能够使利益相互冲突的个体和平共处，共同繁荣。我们在现代社会所需要的机制与道德类似，但却比道德高出一个层次。我们需要一种思维，能够使道德观相互冲突的群体和平共处，共同繁荣。我们需要的是元道德，这种道德体系能够解决不同道德理想之间的冲突；其作用好像群体内部普通的一级道德，能够解决不同个体利益之间的冲突。

元道德并不是全新的概念。相反，早在启蒙运动时期，寻找通用道德规则就是伦理学家们梦寐以求的事情。在我看来，伦理学界长期以来一直试图寻找让人感觉舒服的通用道德规则，但其实这样的规则可能根本就不存在。让人感觉舒服的规则也许只适用于较低层次（群体内部），而在较高的层次（群体之间）则不再适用。也就是说，常识道德或许能让我们避免“公地悲剧”，但对于“常识道德悲剧”，它大概是无能为力的。如果新草地上的牧民想要过上平静快乐的生活，他们也许不得不采取全新的、令人不舒服的方式思考问题。

为了找到我们所需的元道德，我们必须首先理解基本的道德，即能够避免“公地悲剧”的道德。

# 第2章 道德机制

前文提到，道德是一种能够促进合作、避免“公地悲剧”的机制。事实上，道德是很多不同机制的组合，是各种心理能力和性格的组合，这种组合共同作用，促进并维护合作。在本章中，我们将探讨这些机制在心理层面如何起作用，它们与人类的道德思维怎样结合。当然，我们真正想弄懂的问题是：人和人之间为什么会发生冲突？为什么我们的道德系统在新草地上会分崩离析？要想弄懂我们为何会被自己的道德系统背叛（下一章的内容），就要首先了解道德系统在正常情况下的作用原理。

哈丁在公地寓言中描述了一个多人合作的困境。本章当中，我们将问题简化为两人合作困境，就另外一个著名情景——“囚徒困境”展开讨论。尽管这个情景分析的是两位不愿坐牢的囚徒的思维活动，但“囚徒困境”中反映的抽象原则依然能够向我们揭示道德思维的本质。

## 神奇角落

在这个版本的“囚徒困境”中，我们依然给两人起名为亚特和巴德，这次两人结伴去抢银行，他们差一点就成功了，但最终还是不幸被警察捉到，带回警局进行审问。警察确信两人有罪，却苦于没有充足的证据。但因为两人逃税的罪证充足，警察依然能够以这个较轻的罪名，使两人在监狱中度过两年。不过警察最想要的还是以抢劫银行的罪名为两人定罪，这样每人至少会被判处8年监禁。为了给两名囚犯定罪，警察需要获得口供，于是警察把两人分开关押，开始审问。

亚特和巴德面临同样的选择：坦白或沉默。如果两人同时坦白，则各判刑8年；如果两人都保持沉默，则各判刑两年。两人不同的选择和导致的结果以支付矩阵的形式列出（图2.1）

巴德

神奇角落

| | | 沉默（合作） | 坦白（背叛） |
| --- | --- | --- | --- |
| 亚特 | 沉默（合作） | 2年 / 2年 | 10年 / 1年 |
| | 坦白（背叛） | 1年 / 10年 | 8年 / 8年 |

图2.1 经典囚徒困境的支付矩阵

总体来说，两人都保持沉默（合作）时结果更好，但从个人的角度看，选择坦白（背叛）会获得更好的结果。

矩阵中的4个方格分别显示了4种可能的情况。亚特的选择在横向显示，巴德的选择在纵向显示。如果亚特坦白，巴德沉默，则对应左下角的方格，情况对亚特有利，对巴德不利。如果巴德坦白，亚特沉默，则对应右上角的方格，情况对巴德有利，对亚特不利。如果两人同时坦白，则对应右下角的方格，情况对两人都非常不利。而如果两人都保持沉默，对应的是左上角的方格，也就是那个神奇角落，情况对两人都十分有利，两人的总监禁时间也最少。

亚特和巴德将如何选择？你可能会说，两人会一起保持沉默，让自己处于神奇角落中。但如果其他条件不变，两人都从自私的角度考虑，那么这种情况便不会发生。两人都会向警察坦白，就像右下角方格里呈现的那样，每人被判刑8年，是所有情况里两人总监禁时间最长的情况。这个场景与

哈丁的例子遵循了同样的逻辑规律，结果也是同样的“悲剧”。仔细观察图2.1中的支付矩阵便会发现，不论巴德如何选择，坦白对亚特来说都是更好的选择；对于巴德来说也是如此。因此，如果两人都自私且理智，两人便都会坦白。这对警察来说是个好消息，却是两个人的悲剧之源。

囚徒困境与公地悲剧一样，都反映了个人利益和集体利益之间的冲突。从个人角度看，坦白对亚特和巴德都是更好的选择；但从整体来看，共同保持沉默才是更好的选择。我们的问题是：亚特和巴德两人如何才能进入神奇角落？他们如何才能抵御人性的自私，享受到合作的成果？从人性的层面，我们又应当如何自处？在此之前，我们需要对道德的机制进行探讨。

## 家庭观

《犹太法典》中有一个著名的段落：一位对犹太教心存疑虑的人找到犹太拉比希列，他说只要这位伟大的拉比能够在他单脚站立的时间里将整本《托拉》（犹太教的经书）传授给他，他就发誓皈依犹太教。拉比希列回答道：“己所不欲，勿施于人。这就是全部的教义。《托拉》的其余部分都是这句话的注解。仔细领悟去吧。”

当然，所有的主流宗教和我们所知道的伦理学理论都以这样或那样的形式定义了类似的“金科玉律”，这只是其中的一个版本。要想解决亚特和巴德的合作问题，这句话揭示了最直接的方式（并非巧合）。10年的狱中生活是亚特和巴德所“不欲”的情况，如果两人能够采纳拉比希列的建议，就能够共同保持沉默，进入神奇角落。（当然，如果两人真的能够采纳拉比希列的建议，最初就不会选择抢劫银行，不过这就是另外一个问题了。）

亚特和巴德为什么要在意自己是否将“不欲”施于对方呢？也许他们是兄弟，这样一来，问题就解释通了。但若果真如此，我们便又有了新的问题：兄弟之间为什么要彼此关心？著名的亲缘选择理论*从基因的角度看待行为，可以对兄弟之情（或是更广泛的亲情）做出解释。根据定义，有亲缘关系的个体享有部分相同的基因。因此，当某个个体帮助与自己有亲缘关系的另一个个体存活时，便相当于是在扩大自身基因的存活概率。换句话说，从基因的角度看，乐于帮助亲属的基因能够提高自身的存活率，帮助同样优秀的基因片段在他人体内延续。

对很多物种来说，生物学层面的“关心”指的是牺牲自身利益，成全其他个体的利益；并不包含心理学层面的关心。例如，蚂蚁会为同族的亲属提供帮助，但据我们所知，蚂蚁不会被柔情所打动。当然，在人类之间，关心的举动是真情流露的结果，包括我们与骨肉至亲之间强烈的感情纽带。亲情不只是模糊的温暖，也是生物学上的一种策略，是道德系统的一部分，使血缘上有关联的个体共同收获合作果实。

## 投桃报李

亲情能够使血缘上的亲属进入神奇角落，但非亲非故的人们该怎么办呢？如果能够为彼此提供恰当的激励，即使非亲非故，人们也能够进入神奇角落。

假设亚特和巴德对彼此毫不关心，但他们在工作时合作非常默契。若两人合作抢劫银行，结果远远好于两人单独行动或各自与别人搭档。如果他们能够确定最近的一次银行抢劫是他们最后一次合作，那么两人都有理由出卖彼此。但如果两人合作抢劫银行的前景无限，也就是说，只要两人能够抵御诱惑，不向警察坦白，就能迎来美好的未来，两人应当如何选择？如果囚徒困境不是一个独立的事件，而是一系列事件中的某次状况，那么游戏的规则又会发生变化。亚特当然可以在背后出卖巴德，为自己赢得短暂的一年监禁，让巴德面对长期监禁。但这样一来，亚特就等于放弃了与巴德合作的美好未来，为了少在监狱里待一年而放弃这样的未来似乎并不值得。因此，如果亚特和巴德把眼光放长远，两人就会共同保持沉默。这并不是因为他们彼此关心，而是因为他们对彼此有用，因为两人想在未来获得丰厚回报，当下就必须选择合作。“我帮你抓痒，因为你会帮我挠背。”——这种有条件的合作被称为互惠原则，或互惠利他主义。*

20世纪80年代初，罗伯特·阿克塞尔罗德（robert axelrod）和威廉·汉密尔顿（William Hamilton）发表论文，介绍了一场“囚徒困境锦标赛”的研究结果，被后人奉为经典案例。这场比赛的选手不是真人，而是计算机。研究人员在计算机中设定不同程序，使用不同的策略在囚徒困境中博弈。最简单的策略有两种，一种是无条件合作（总是保持沉默），另一种是绝对不合作（总是选择坦白）。（不合作也被称为典型的“背叛”。）阿克塞尔罗德和汉密尔顿请他们的同事各自编写程序，参加这次比赛。参赛的大部分程序都十分复杂，但最终获胜的程序却非常简单，就像“无条件合作”和“绝对不合作”的策略一样直接。最佳程序由阿纳托尔·拉波波特编写，他的策略被称为“投桃报李”，这个策略首先从选择合作（保持沉默）开始，之后每一轮都采用同伴上一轮的选择。如果上一轮对方选择合作，那么本轮己方也选择合作，反之亦然，“投桃报李”便由此得名。在后续进行的比赛中，的确有其他程序将“投桃报李”策略击败，但这些程序的基本思想依然是“投桃报李”策略。互惠原则是十分实用的。

人类可以通过有意识的推理得出互惠的逻辑：“巴德上次出卖了我，他可能会再次将我出卖，因此，这次我不会试着与他合作。”而巴德当然也可以通过推理预测出亚特的心理：“如果我这次出卖了亚特，他以后就会认为我不愿选择合作。但相较于现在出卖亚特而言，今后与亚特合作会给我带来更多的好处。所以我现在应该与他合作。”这种清晰的战略考量可以

将亚特和巴德引入神奇角落，但这种思考往往是多余的，因为人类的感觉会自动完成这种思考。假设巴德背叛了亚特，亚特可以通过推理决定抛弃巴德。但如果亚特对于巴德的背叛能够自动报以气愤、憎恶或蔑视*的情绪，最终也会达到与推理同样的效果，中途发生意外的可能性也会更小。同样，巴德可能会领悟到，如果他出卖了亚特，亚特将会对他心存恶意，这种情况将会影响巴德的职业生涯。一想到背叛亚特，巴德可能会打个寒战。同样，正面的情感也会通过互惠原则促进合作。通过与亚特合作，巴德可能会期待亚特对自己心存感激，从而更加愿意与自己进行下一步合作。

人类的近亲灵长类动物似乎也进行过有条件的合作。考虑到它们的合作水平，其合作可能更多基于情感，而不是基于清晰的逻辑推理。一个研究黑猩猩分享食物行为的经典案例发现，成年黑猩猩更愿意把食物分给近期为自己梳过毛的黑猩猩。如果近期没有为成年黑猩猩提供过梳毛服务的其他猩猩前来觅食，有食物的一方往往会强烈抗议，做出攻击性动作。类似的研究表明，人类互相帮助、投桃报李的能力取决于，至少部分取决于我们从灵长类祖先那里继承的情绪化性格。

如果互动得当，被动情感能够激励合作性行为；但如果过度应用，被动情感也可能会毁掉合作关系。假设巴德由于一时软弱，向警察坦白了事实，将亚特陷于意料之外的不利境地。多年之后，亚特和巴德发现了百年难得一遇的机会抢劫银行，如果亚特还在生巴德的气，就可能会错过这个机会。宽恕需要代价。（想象几位上了年岁的摇滚乐手，为了丰厚的出场费而尽弃前嫌，共同出席巡回演唱会。）计算机模拟的结果也同样表明，在充满变数的世界里，有条件合作的一方如果更加宽容，最终的结果会优于无休止的怨恨。黑猩猩的行为似乎也印证了这一点。德瓦尔和罗丝马伦观察了上百段黑猩猩之间发生冲突后的录像，他们发现，打完架后，黑猩猩往往会互相亲吻、拥抱。这可以说明，在充满变数的世界里，宽恕能够调节负面的被动情感，是符合互惠原则的行为。人类的宽恕能力在生物学上能够追溯到很久以前。

## 最好的朋友

亚特保持沉默的原因可能是害怕巴德发怒，导致这段回报颇丰的合作关系走向毁灭。但经过多年的合作之后，两人的合作方式可能会发生变化。亚特和巴德可以计算各自的长期成本和收益，以此作为合作的动力，这也是可行的办法。如果两人能够通过直觉感知到这种方式，最终效果可能还会更好。具体说来，对亚特和巴德这样的银行劫匪来说，如果他们能够建立心理机制，下意识地关心可能会与自己合作的人，对他们是有好处的。

那么，这种心理机制如何运作？具体说来，这样的机制如何判断哪些人与自己的合作前景更加美好？历史是未来最好的向导。如果一个人曾经与你合作多次，你们在未来也可能会有更多的合作。因此，如果某种心理机制能够促使一个人关心自己曾经的合作伙伴，那么合作就会变得更加自然、有效。我们可以把这种心理机制称为友谊。

如果将友谊的主要内容看作合作，而不是闲逛、玩耍等其他活动，会让人产生奇怪的感觉。然而，表象是具有欺骗性的。首先，大自然的目的并不一定要在生活中反映出来。比如说，性行为最初的目的是繁衍后代，但人们不一定非要在需要繁衍后代的时候才进行性行为。同样的道理，我们对他人伸出友谊之手时，我们心中所想的可能与友谊最初的目的相去甚远。事实上，如果你总在思考自己从友谊当中得到的实在利益，说明你并不是真正的朋友。其次，如果把友谊视为合作策略的想法让人感到奇怪，这可能因为我们正处于非同寻常的好时代。以渔猎采集为生的人类祖先过着饥一顿、饱一顿的生活，对他们来说，来自朋友的用餐邀请远不只简单示好，而是关乎生死的大事。他们所处的世界远比现在简单粗暴。当今世界，很少有人能够说自己曾经救过朋友的性命，但远古时期的世界却远非如此。最后，很多情况下，合作给人的感觉并不像是“合作”，记住这一点十分重要。朋友之所以成为朋友，不仅是因为大家在一起时共同做的事情，更是因为大家各自分开后不会去做的事情。朋友不会偷你的东西，也不会说你的坏话，更不会与你的爱人偷情。这些日常的友善行为虽然微不足道，但它们都是合作的一种方式，就好像亚特和巴德在小路上相遇时擦肩而过，不发生任何意外一样。这种合作被我们称为“友谊”，它始于善意的亲近，并由此进一步发展。

## 起码的尊重

假设你是亚特，正在寻找一位搭档抢劫银行。你听说有个人名叫巴德，手脚十分利索，开车逃跑的能力也是一流。但巴德唯一的缺点是，只要形势对他有利，即使从二楼用枪打穿你的脑袋他也在所不惜。亚特当然不是理想主义者，但鉴于抢劫银行可能会遇到各种不确定因素，他实在犯不上为此和精神不正常的巴德合作。结论：两位互不相识的人要想合作，双方对彼此的利益都要有起码的尊重。

如前文所述，在风险很小的情况下，雄性黑猩猩倾向于将陌生猩猩杀死。有时候，人类也会将陌生人视作威胁，希望将其除掉，甚至吃掉。（据说南太平洋上的食人族会将可食用的闯入者称为“人肉”。）但不论怎样，人类都可以选择对陌生人减少一分恶意，事实上，我们在当今世界也正是这样做的。美国南北战争时期，很多北方士兵在战场牺牲，但他们手中的枪支却一枪未发，此类报告十分频繁，北方军队的高级官员曾一度感到十分

沮丧。许多士兵实在无法说服自己向陌生人开枪，即使那些陌生人想要杀死他们。美国军方由此得出结论，士兵需要经过训练，克服对杀戮的抗拒。现代的军事训练便由此而来。

近期，菲耶里·库什曼、温迪·门德斯和同事们在实验室进行了一项试验，研究人类对暴力行为的抗拒心理。他们让参试者模拟一系列的暴力行为，比如用锤子击打他人的腿部（如图2.2），同时监控参试者的生命体征。

所有的参试者都十分清楚，自己的行为不会造成任何伤害。然而，仅仅是摆出令人不快的动作，他们的外周血管就会剧烈收缩，反映出“胆战心惊”、手脚冰凉的生命体征数据。此外，研究人员还发现，参试者只有在亲手模拟暴力行为时才会出现血管收缩的反应。而观看他人模拟暴力行为，或自己做出姿势相仿的非暴力行为（比如用锤子钉钉子）时，其生理体征并不会出现如此剧烈的变化。很多参试者对暴力行为的模拟都尽量敷衍了事，比如只用锤子在假想受害者的腿上轻轻敲一下。还有一位参试者拒绝做出这个动作。当然，人类有时会变得非常暴力，而且诱因往往微不足道。但即使暴戾如斯，也已经是打了折扣的行为。通常情况下，仅仅在想象中对无辜者甚至陌生人暴力相向，就会令我们不寒而栗。这也许是人类道德思维中至关重要的一个特点。（试着想象没有这种机制的世界会是怎样。）



图2.2 参试者模拟暴力行为时，虽然知道自己的行为不会造成任何伤害，但他们的生理体征仍然反应强烈。标注X的是一条假腿

除了对善意行为的友好回馈之外，起码的尊重还包含更多的内容。人类本可以比现在更加善良，我们现在至少愿意为他人，甚至是陌生人行些方便，不求回报。20世纪60年代，斯坦利·米尔格拉姆（stanley Milgram）和他的同事进行了一项经典试验。他们将“丢失的”信件放到公共场所，很多信件甚至连邮费都没有付，但大多数信件依然被寄了回来。即使明知不会再次光临，我们依然会在饭店留下小费。还有些人会向慈善机构匿名捐款。有一种观点认为：人类之所以会为他人提供帮助，往往是由于我们为他人感到难过，希望能够减轻他人的痛苦。大部分人对此半信半疑，但研究人员始终持质疑态度。社会与发展心理学经过几十年的研究，终于证实了这个观点。在囚徒困境试验中，研究人员用金钱的输赢代替坐牢的惩罚，经研究发现，为他人感到难过会使人更乐于选择合作。（后文会有大量篇幅研究以金钱为筹码的合作游戏。）这种替别人感到难过的心理被称为同理心，指的是对他人的感情感同身受的心理状态。*近年来，认知神

经学家对同理心的神经机制进行研究，发现同理心这个名称实在恰如其分：人类目睹他人经历痛苦时所触发的神经回路与人类亲历痛苦时所触发的神经回路完全相同。同理心强的人，其大脑中类似的活动则更加明显。

对陌生人产生同理心所牵涉的神经回路最初可能由负责怀孕生产的神经回路发展而来。催产素是一种神经递质，是哺乳动物怀孕生产过程中一种十分重要的激素。控制大脑对催产素敏感程度的基因与同理心强弱紧密相关。人工向鼻腔内喷洒催产素（催产素能够由鼻腔进入大脑）也能够使人在囚徒困境中倾向于合作。

人类关心他人（包括没有亲缘关系的陌生人）的能力充分显示了我们从灵长类祖先那里继承来的特质。过去几十年的研究中，灵长类动物学家多次发现猩猩和猴子做出怀有怜悯之心的举动。该领域科学家娜杰日达·拉德金娜·科茨（Nadezhada ladygina-kohts）十分具有探索精神，她在莫斯科的家中饲养了一只黑猩猩，取名为乔尼。乔尼爱在屋顶玩耍，不愿下来。随着时间推移，科茨发现让乔尼从屋顶下来的最好方法就是激发它的同情心。如果科茨假装哭泣，乔尼就会迅速跑到她身边，警惕地四处张望，寻找欺负科茨的人，还会轻轻触碰科茨的脸，表示安慰。有时候，黑猩猩也表现出乐于助人的行为。荷兰阿纳姆动物园里有一只7岁的黑猩猩，名叫杰西。一位上了年岁的管理员克罗姆试图取回一只灌满水的轮胎，几次失败后，克罗姆沮丧地放弃了尝试。杰西看到后，便跑到轮胎旁边，移开了挡路的其他轮胎，小心地将这只灌满水的轮胎搬到了克罗姆身边，水一点都没有洒出来。

这类小故事十分有趣，也能反映出人类这个灵长类近亲的一些本性。但如果有人对这些故事仍存有疑惑，我们可以进一步解释清楚。最近，灵长类动物学家在实验室设计了对照试验，印证了非人类灵长动物互相关心的行为。通过一系列试验，菲利克斯·瓦纳肯（Felix Warneken）、迈克尔·托马塞洛（Michael Tomasello）和他们的同事证明，黑猩猩会自发地对人类和其他猩猩提供帮助，并且不期望得到任何回报。一次试验中，黑猩猩主动帮助试验员取回远处的物体。另一次试验中，黑猩猩爬过障碍，为一位陌生人提供帮助。还有一次试验，黑猩猩主动为另一只猩猩解开锁链，让它能够吃到食物，但自己却没有得到任何的好处。睦邻友好这种想法似乎由来已久，甚至能够追溯到进化史更早期的阶段。文卡特·拉克斯铭亚南（Venkat lakshminarayanan）和劳里·桑托斯（laurie santos）近期的试验中，僧帽猴面临两个选择，一是只给自己奖励，二是奖励自己之外，再奖励一位邻居。大部分僧帽猴都会选择与邻居一起获得奖励，即使邻居获得的奖励胜过自己也没关系。同理心甚至在老鼠身上也有反映，它们甘愿放弃近在眼前的奖励，而选择让另一只被困的老鼠重获自由。

总之，人类是懂得关心他人的物种，并且对他人的关心会有一定的限度。

这种能力部分来源于我们的灵长类祖先，甚至是更久远以前的祖先。我们最关心的自然是亲属和朋友，对熟人甚至陌生人也同样会表示关心。一般情况下，我们极其不愿伤害陌生人，即便只是假装做出伤害的动作，我们的血管也会收缩。同时，在不做出太多牺牲的前提下，我们还会不计回报地帮助陌生人。因为我们彼此关心，关注个人利益以外的内容，所以对我们来说，进入神奇角落并没有那么难。

## 威胁与承诺

如果亚特和巴德相互关心，或者拥有辉煌的合作前景，两人便能进入神奇角落。但倘若两人素不相识，未来也没有交集，那么又当如何？假设亚特和巴德面临一次千载难逢的好机会可以抢劫银行，两人之前从未进行合作，以后也不会再次合作。毫无疑问，警察逮捕两人后会试着策反他们。这种情况下，亚特和巴德还会互相包庇吗？

也许两人可以事先约定，共同保持沉默。做出这样的承诺并不困难，难的是遵守承诺。问题在于，做出承诺本身并不会影响支付矩阵的结果。当坦白还是沉默的选择真的来临时，对亚特最有利的选择依然是坦白，对巴德也是如此。如果两人并不关心彼此，也不必关心两人未来的合作，那么不论有没有承诺，两人都会选择坦白。

因此两人需要一种方式巩固约定。为了达到这样的目的，亚特可能会对巴德说："如果你出卖了我，我出狱后就会找你算账，将你杀死。"不幸的是，威胁策略与更加友善的承诺策略有着同样的问题：假设亚特做出了威胁，而巴德依然出卖了他，10年之后，两人都被释出狱时，亚特便可以将自己的威胁付诸实践。但那时的亚特为什么还要费力报复呢？杀掉一个人是有风险的，并且亚特从中无法得到任何好处。可是如果巴德从一开始就知道亚特不会费力报复，那么亚特的威胁就是徒劳的。巴德会忽视亚特的威胁，选择坦白。当然，如果两人位置对换，亚特也会做出同样的选择。这种情况下，两人无法合作。

可以看出，单纯的威胁是无效的，单纯的承诺也同样无效。但如果方法合适，威胁便可能产生效果。假设亚特拥有一个高级机器人杀手，可以通过编程控制。如果巴德出卖了亚特，亚特可以编写程序，让机器人追杀巴德。重要的是，我们假设亚特的机器人完美无缺，一旦程序启动，便再也无法停止，即使是亚特本身也无能为力。如果巴德知道一旦自己背叛亚特，就会必死无疑，那么他一定不会选择背叛。亚特的威胁有些疯狂，因为他必须对机器人的行为负责，他当然不愿看到自己的威胁变成现实。如果巴德出于某些原因忽视亚特的威胁，出卖亚特，亚特也会全力阻止自己的机器人，即使这种努力徒劳无功。尽管如此，通过事先设立疯狂的威

胁，亚特便能够在巴德充分了解情况，并且头脑清醒的前提下，确保巴德与自己合作。当然，巴德也可以通过同样的方式，确保亚特与自己合作。（这个策略可以叫作“同归于尽”。）

可惜呀，人类尚未发明出通过编程操作的机器人杀手。但经济学家罗伯特·弗兰克（robert Frank）的研究显示，人类大脑内有一种情感机制能够起到相同的作用。*假设亚特性急鲁莽，一旦遭到背叛，亚特将会暴怒，不顾一切杀死巴德，即使要等10年时间，即使要追巴德到天涯海角，他也在所不惜。如果巴德知道亚特如此有仇必报，那么巴德很可能不会选择出卖亚特。因此，如果亚特有仇必报，并且因此闻名，亚特就可以充当自己的机器人杀手，通过严重、可信的威胁，促使他人与自己合作。当然，有仇必报是要付出代价的。如果亚特真的将毕生精力用于报复巴德，最终可能会使自己一无所有。但是如果不出意外，亚特永远也不必追杀巴德，因为巴德和其他人并没有胆量出卖亚特。因此，促使人们有仇必报的情感可以是一种理性的非理性。这种情感促使我们公开承诺，承诺自己将会做出对自己不利的事情，因而从中受益。

复仇的天性并不是人类所独有的。基思·詹森（Keith Jensen）和几位同事设计了一个试验，在试验中，黑猩猩可以通过拉绳子来破坏放食物的桌子，从而阻止其他猩猩获取食物。研究发现，如果a猩猩从b猩猩处偷取了食物，b猩猩就更可能拉动绳子，使a猩猩无法够到食物。野外研究显示，生活在野外的黑猩猩也会做出类似的行为。

负面的社会情感能够促成我们与他人的合作，但更加高尚的情感同样也可能促成合作。如果亚特和巴德是两个无赖，那么他们的承诺毫无意义。因为如前文所述，无赖不会遵守承诺，大家对这一点心知肚明。但如果亚特和巴德是高尚的小偷呢？亚特也许不关心巴德，但他十分关心自己是否能够履行承诺。如果亚特失信于人，他可能会对自己失望至极，恨不得立即把自己扔到最近的熔岩坑里。那么，这样的人值得我们与之合作。复仇的怒火能够使我们做出更加可信的威胁，而易于产生强烈的内疚和羞愧则是我们对自己造成的威胁。如我们所料，真的违背自己的诺言，或者只是简单的想象，就会使大脑内部控制情感的区域活跃起来。

前文提到，亲情和友情都是合作性关心的表现形式。同复仇的怒火一样，类似的情感可以作为战略约束，让我们理性地承诺做出不理性的行为。然而，这种情况下，受到情感约束的双方并非再无合作的可能，只是如果他们选择与别人合作，未来可能会更加美好。假设警察给亚特开出了诱人的条件：如果亚特出卖巴德，警察不仅可以让亚特全身而退，而且还会给他找一份工作，作为银行抢劫专业顾问。也就是说，警方邀请亚特与他们，而不是巴德进行合作。亚特关心巴德，是因为两人过去曾经合作，而且未来的合作前景十分美好。但警察给亚特开出的条件是巴德过去只能在梦中

想象的——刺激且有地位的工作和稳定的高额收入。亚特与警察的合作前景似乎已经超过了与巴德的合作。如果亚特与巴德的友情建立在两人的合作前景之上，前景越好，友谊越深，那么亚特将会抛弃巴德，与警察合作。择优交易对亚特是有利的，但如果亚特在人们心中的形象变成了择优交易者，就不是什么好事了。因为巴德同其他人一样，也许根本不会与见利忘义的人进行合作。

这就涉及了忠诚这种美德。如果亚特把抢银行的同伴看得比其“市场价值”（与他们合作抢劫所产生的价值）更重要，那么亚特就是一位有魅力的合伙人。史蒂芬·平克认为，忠诚在恋爱关系中能够得到最好的反映：你的条件十分优秀，但总会出现一个人，不仅拥有你全部的优点，还比你略微好一点点。虽然你知道自己的爱人某天可能会遇到这样的人，但你也知道，你的爱人不会喜新厌旧，将你抛弃。这种信心会让你更加愿意与爱人安定下来，建立家庭。如果将这个过程视为合作，那么这就是高风险的合作。你有很多物质上的优点，你的爱人对你十分欣赏，这样当然很好，但两个人要想永不分离，这一点也许还不够。你真正想要的，是爱人发自心底、不可动摇的爱，是只想和你一个人在一起的愿望。简而言之，就是你希望爱人能够爱你，不仅因为你拥有诸多优点，而是因为你就是你。只有爱才能让你足够忠诚，甘心承担起为人父母的重任。因此，爱不仅仅是强烈的关心，它是一种极其特殊的心理机制，是一种情感约束。它使父母双方确信他们不会抛弃彼此，从而达成结婚生子的合作。

还有一种忠诚也同样能够推动合作的车轮。对个人的忠诚能够使人成为更好的朋友、更好的爱人；在更大的合作团体中，对权威的尊重能够使人成为更合格的小兵。如果你是一名陆军上将或是执行总裁，你想要怎样的人加入团队？忽略你的意见、自行其是的人，还是服从指挥、踏实可靠的人？同样，你想要的是一看到更好的选择就弃你而去的人，还是一直坚守岗位，直到最后一刻的人？好的小兵忠诚且谦逊，他们安分守己，不敢背弃职责。

积极情感和消极情感都会使人安分守己。几乎所有的灵长类动物中，地位较低的个体都对地位较高的个体怀有消极情感，恐惧之心占了多数。但人类提到自己的领袖时，有时会怀有强烈的敬佩之情。我们可能被从未谋面的领袖所激励，投身加入某些成员并不固定的组织，比如国家、教堂、公司、学校等。乔纳森·海特（Jonathan Haidt）认为，对领袖、组织，以及更加抽象的理想表示忠诚的能力，最初可能是为了促进大规模团体内部的合作发展而来，就好像爱情最初是为了促进父母合作，养育子女一样。这种表示忠诚的能力也许取决于人类敬畏的能力，也就是看到比自己和自己熟悉的社交环境更加宏大的事物时，为之感动并献身于此的能力。

## 警醒的眼睛与洞察一切的心

亚特和巴德要想进入神奇角落，两人就需要互相关心，或是考虑到两人未来的合作，或是两人在情感上被威胁或承诺所约束。但如果两人之间一无所有，又当如何？

亚特曾经发誓，如果巴德背叛自己，就将他杀死。但巴德独自一人待在牢房，依然会犹豫是否坦白。巴德之所以犹豫，是因为他知道亚特是理性的。亚特并不鲁莽，他不会仅仅因为愤怒就去追杀巴德。然而，巴德依然需要三思，亚特可能会因为巴德的背叛而将他杀死，这并非因为亚特有仇必报，而是因为其他人都在观察。银行抢劫界想要知道的是，亚特的威胁是否可信。如果亚特因为巴德的背叛而将他杀死，那么答案就是肯定的，这对亚特来说是件好事。因此巴德最好保持沉默。

由此可见，有名声负累的人倾向于选择合作。声誉促使人们兑现威胁，使自己的威胁变得可信，也促使被威胁一方选择合作。此外，声誉也会更加直接地推动合作。如果巴德因为背叛而臭名昭著，那么就不会再有人愿意与他合作，巴德终将成为输家。因此，声誉对合作的推动有两种方式：促使人们证明自己的合作愿望，或是促使人们证明自己拒绝容忍背叛。这两种方式都符合人类道德思维的本性。

凯文·海利（Kevin Haley）和丹尼尔·费斯勒（Daniel Fessler）设计了一个试验。他们选出一半参试者，交给每人10美元，然后给拿到钱的参试者一个机会，由他们选择是否与没有拿到钱的参试者分享这10美元，或者干脆把10美元全部送出。这个试验被称为“独裁者博弈”，因为被选中的一方能够控制全部资金。整个试验以匿名方式在互相联网的电脑上进行，确保参试者对他人的选择毫不知情。试验中的关键变量十分隐蔽：在被选中的“独裁者”中，一半人的电脑以一双眼睛的图案作为桌面，如图2.3所示；另一半“独裁者”则只能看到标准的桌面背景，上面画了实验室的标识。




图2.3 海利和费斯勒的试验中使用的眼睛图案。看到这双警醒的眼睛，人们对陌生人表现得更加慷慨

看到标准桌面背景的人只有一半左右（55%）选择分享。而在桌面背景上看到眼睛的人绝大多数（88%）都选择了分享。在随后进行的街头试验中，研究人员使用“无人售货机”出售饮料，结果显示，画有眼睛的图片会使人付出两倍的价格购买牛奶。

众所周知，如果人们认为别人在观察自己，就会变得更加自觉，更加注意自己的行为。这个试验的奇妙之处在于，一个无关紧要的低级暗示——一张画着眼睛的图片，就能将人们最好的表现激发出来。说它“无关紧要”，是因为没人会有意识地对这个暗示做出回应，没人会认为“因为橱柜上贴了一张画着眼睛的图片，所以我拿了牛奶需要付钱”。这是一种自发的程序，是道德机制中一个有效的环节。

警醒的眼睛会对我们产生巨大的影响，也许是因为眼睛背后往往有一张发表评论的嘴。人类学家罗宾·邓巴（robin Dunbar）指出，人类的对话中，65%的时间都在议论他人或好或坏的行为，对他人说短道长。他认为，人类之所以花大量时间议论他人，是因为议论能够促进合作，是一种重要的社会制约机制。事实上，一想到“所有人”都会知晓自己的所作所为，大部

分人都会选择谨慎行事。况且，议论他人不仅是人类的一种能力，更是一种自发的行为。对很多人来说，不对他人说短道长是很困难的。

在警醒的眼睛和爱说闲话的嘴巴遍布周围的世界里，如果人们做了不利于合作的事情，迟早会被人发现。对于一个被抓了现行的背叛者来说，最坏的情况可能会不堪设想：在余生当中，没人愿意与你再有任何瓜葛。那么如何才能避免这种命运？如果能够让“所有人”相信你以后会努力合作，那么事情还有转机。你可以道歉，但这显然不够，因为所有人都会说“对不起”。如果你的脸部不自觉地浮现出不自然的颜色（比如涨得通红），这种道歉可能会更有说服力。因为这表示你真心地为自己的行为感到羞愧。事实上，这也许就是窘迫这种情感产生的原因：通过表达真心悔改的诚意，帮助人们重拾社会地位。这种方式似乎是有用的，研究表明，如果背叛者在事后表现出窘迫的样子，便更容易取得人们的原谅。

当然，如果“所有人”的态度仅限于知道某人的背叛行为，这本身并没什么。关键在于，人们会基于自己看到和听到的事实，改变对一个人的态度。在警醒的眼睛和竖着的耳朵背后，让我们心存顾虑的其实是人们论断是非的心，人类会论断是非，这早已不是新鲜事。但值得注意的是，早在婴儿时期，人类就已经开始分辨是非。10年前，一个著名的心理学试验证明了这一点。

基利·哈姆林（Kiley Hamlin）、凯伦·韦恩（Karen Wynn）以及保罗·布鲁姆（Paul bloom）给6个月和10个月大的婴儿播放动画，画面中有眼睛圆圆的几何体人物，可以沿山坡上下滑动。如图2.4所示。



图2.4 还不会说话的婴儿喜欢小三角，因为它帮助圆形爬上了山顶；不喜欢小方块，因为它把圆形推下了山

在左图所示的动画中，圆形想要爬上小山，但总是无法爬到山顶。于是小三角从山脚爬上小山，前来帮忙，把圆形推到了山顶。在右图所示的动画中，圆形依然想要爬上小山，却无法到达山顶。随后出现了前来捣乱的小方块，它从山顶下来，把圆形推回了山脚。试验人员为婴儿们反复播放两段动画，直到他们感到厌烦。接下来关键的试验阶段中，研究人员给婴儿

们拿来了一个托盘。托盘里一边放有像是小三角的玩具，另一边放着像是小方块的玩具。*16名10个月大的婴儿中，有14名选择了动画中前来帮忙的小三角，而12名6个月大的婴儿则全部选择小三角作为玩具。如此显著的试验结果着实让人惊讶。

接下来，研究人员在另外一组婴儿中重复了这项试验。但圆形上不再有圆圆的眼睛，没有了拟人的效果；前期圆形试图自己滚上山坡的片段也不再播放，消除了圆形行为的目的性。因此在这次试验中，圆形不再是想要达成目标的个体，而像是一个没有生命的物体。此外，小三角和小方块也不再以帮忙者或捣乱者的身份出现，它们只是简单地把圆形推上或推下山坡。在这一次选择玩具时，婴儿的行为与研究人员的预期相一致，他们对小三角和小方块（向上推的图形和向下推的图形）没有了特殊的偏好。这说明，婴儿的偏好是社会化的：他们偏爱的是帮忙，不是向上推；嫌恶的是捣乱，而不是向下推。

6个月大的婴儿远未学会走路和说话，但他们已经开始对行为和个体进行价值判断。他们愿意接近有合作倾向的（关心他人）个体，回避不愿合作的个体。因为这些婴儿年龄尚小，他们的行为显然不是理性推理的结果。他们不会想到“小方块对红色的圆形不好，说明小方块可能也会对我不好，所以我要回避小方块”。相反，婴儿的判断由自发的程序完成，这种程序对于低级暗示十分敏感，注意到了某些运动的含义和看上去像是眼睛的东西。鉴于这种机制出现得如此之早，我们几乎可以断定，这是人类通过基因遗传下来的能力。

## 外人免入

两位囚徒，如果他们彼此关心，或者在未来拥有合作前景，就能够进入神奇角落。两位陌生人，如果他们恰当地使用威胁，或是两人为声名所累，也能够进入神奇角落。但是在没有威胁、没有声名之累的情况下，陌生人之间可能建立合作关系吗？

假设有一群银行劫匪，成立了“守口如瓶银行劫匪联盟”。顾名思义，联盟中的劫匪面对权威时，都会遵守严格的沉默守则。联盟的规模很大，大部分成员彼此都互不相识，既没有私人交往，也不曾听闻他人的名字。换句话说，联盟里的成员都是陌生人。但加入联盟就意味着丰厚的收获，尽管素不相识，联盟成员依然可以一同抢劫银行，并相信同伙不会出卖自己。问题解决了吗？

这个假设好像在用制定规则的方式解决问题。假设这样一个联盟的存在，就是规定了一个合作群体的存在。问题在于如何使联盟开始运行，如何防止联盟解散。加入联盟意味着做出承诺，承诺自己不会背叛其他成员，以

此保证自己也不会被他人出卖。但如前文所述，在自私的世界中，单纯的承诺毫无用处。如果违背诺言不必付出任何代价，联盟成员为什么还要遵守承诺呢？也许，违背诺言的人会受到联盟的惩罚。这确实是个办法，但又引发了下一个问题：联盟由谁来管理？管理者对陌生人之间的背叛实施惩罚，又能得到什么好处？实施惩罚的问题可以稍后讨论。现在我们可以将情景简化，假设联盟的成员生来就守口如瓶。只要他们的合作仅限于联盟内部，所有的成员就是安全的。联盟成员需要注意的是，避免口风不紧的联盟外成员参与行动，对联盟成员加以利用。劫匪无赖可能会设法靠近一群守口如瓶的银行劫匪，高兴地与联盟成员多次合作，每次都在警察询问口供时神不知鬼不觉地将自己的同伙送进监狱。

如果劫匪中的坏人声名狼藉，联盟成员就能够预先知情，避免与劫匪无赖合作。但根据我们的假设，这种情况不可能发生。既然无法掌握不可信的外人的信息，联盟成员可以主动提供可信的自己人的信息。联盟可以给每位成员发一张身份卡片，上面盖一个“守口如瓶章”。这样，联盟成员就能够互相辨认身份，避免与外人合作。只要外人无法伪造身份卡片，这样的身份认证系统就是可行的。为了保证联盟运作正常，就需要一套可靠的成员加盟标准。

这是一个常见的问题：所有的合作组织都需要保护自己成员的利益。这就需要人们有能力区分“我们”和“他们”，并且将“我们”的利益置于“他们”之上。尽管极少数人会将陌生人与家人一视同仁，但所有的人类社会都不会将这种规则作为社会规范。究其原因，如果按照这种规则运行，社会将会变成开放的资源供应站，将财富无私地散播给每一位陌生人，即使是初来乍到，享受的待遇也堪比失散多年而后重逢的亲人。人类学家唐纳德·布朗（Donald brown）通过研究文化间的差异与共性，也证明了群体对内部成员的优待和民族优越感的普遍存在。

以每个人分别作为圆心，能够画出很多代表社交圈子的同心圆。离我们最近的小圈里，是我们最亲密的家人和朋友，外围的大圆则代表关系略远的亲戚和熟人。在亲友圈之外，是我们所属的团体中与我们有关的陌生人。这些团体大小不同，种类各异，包括村庄、宗族、部落，街坊、城市、州、地区、国家，教会、教派、宗教。除了这些由小到大排列的组织之外，我们还根据政治立场、就读的学校、社会阶级、支持的球队，以及其他的喜好和厌恶为自己划分组织。社会空间是复杂的，包含多个维度。但不论是基于常识还是基于大量的社会科学研究，至少有一点可以确定：在以自己为圆心的社交网络中，人类对他人在圈中所处的位置极其关心，倾向于优待距离自己更近的人。我们将这种倾向称作部族主义，有时也叫作狭隘的利他主义。

将社交圈最内层的人（家人、朋友、熟人）视为合作对象并不困难，但人

类的合作范围往往更加广泛，有时是积极的合作，比如合作修桥、战争中的合作；有时则是更为消极的合作，比如表现出友好的态度。然而，与陌生人合作的前提是制定标准，辨别哪些陌生人可以与之合作，而哪些陌生人可能会利用我们。换句话说，我们要有能力颁发并辨认社会身份卡片，并且能够根据辨认结果调整自身行为。

《希伯来圣经》中记叙了基列人的故事：大约公元前1200年，基列人击败了以法莲人，驱逐他们离开故土，渡过约旦河。这场战争后，很多幸存的以法莲人试图瞒过基列人在河边的守卫，回归故里。为了拦截以法莲流亡者，基列守卫采用了一个简单的测试：想要通过关卡的旅人必须朗读一个希伯来单词：shibboleth（植物结实的部分）。古代以法莲方言中没有sh的发音，因此以法莲人很难读准这个单词。据《圣经》记载，因为不会发sh的音，有4.2万名以法莲人因此丧命。

现今，shibboleth的意思是，能够辨别文化群体内部成员身份的可靠印记。凯瑟琳·克林泽（Katherine Kinzler）与其同事的研究表明，在发展早期，人类就倾向于使用语言这种原始的“印记”来标记群体身份，以此作为社会偏好的基础。在一组连续试验中，研究人员选取英国和法国儿童作为研究对象。他们发现，6个月大的婴儿更喜欢将目光投向说话不带外国口音的讲话人；10个月大的婴儿更愿意从说本国语言的人手中接受玩具；5岁大的孩子更愿意和没有外国口音的孩子做朋友。似乎早在习得语言能力之前，大脑就已经开始用语言来分辨可靠的“我们”和不可靠的“他们”了。

语言印记揭示了部族主义中更加宏观的一个方面：随机差异造成的影响可能并不随机。基列人如何发音本身并不重要，重要的是基列人的发音与以法莲人不同。同样道理，随机的文化行为在促进合作方面可能会起到重要作用。人们着装、清洗、饮食、工作、舞蹈、唱歌、开玩笑、约会、做爱的方式，以及日常生活中遵循的所有规则，都可能造成并不随机的影响，让陌生人显得突兀，将“我们”与“他们”区分开来。

当今世界，“我们”和“他们”之间最为明显的一个界线就是种族。近年来，心理学家利用内隐联想测验（iaT）对不同种族人群的观点进行研究。内隐联想测验通过衡量人们归类划分不同概念的速度，总结出概念间的联系。（你自己也可以试试）*在典型的内隐联想测验中，参试者需要在电脑上完成两个内容交错的分类任务。例如，参试者会被要求将屏幕上出现的词语分类，判断它们指代的含义是好还是坏。一个人看到好的词语（如：爱）时可能会按下左手的按钮，看到坏的词语（如：恨）时会按下右手的按钮。同一个人看到不同的头像照片时，便可能会根据头像的种族分类，比如看到白人就按下左手的按钮，看到黑人就按下右手的按钮。内隐联想测验测量的是人们给词语分类的速度，以及不同分类对应不同按钮时人们判断速度的变化。例如，看到好的词语和白人照片时，你可能会按

下右手的按钮，而在看到坏的词汇和黑人照片时按下左手的按钮。但有时，你可能会在看到坏的词汇和白人照片时按下左手的按钮，看到好的词汇和黑人照片时按下右手的按钮。如果你使用同一个按钮表达“坏的”和“黑人”时所用的判断时间更短，那就说明你的潜意识里认为“坏的”和“黑人”之间存在某种联系。对于其他的概念搭配，我们也可以做出相似的推断。内隐联想测验结果显示，多数白人潜意识中都更加偏爱白人，他们更容易将好的词汇与白人头像归为一组，而将坏的词汇与黑人头像归为一组。大脑的活动情况也能够反映出内隐联想测验的分数：在黑人头像与“坏的”之间建立更强联系的人，看到黑人头像时，大脑中与提高警惕性相关的区域（杏仁核）也会做出更加强烈的神经反应。为儿童设计的一项内隐联想测验发现，儿童在年仅6岁的时候就已经显示出种族偏见，与成年人的种族偏见如出一辙。令人吃惊的是，一项为猴子设计的内隐联想测验表明，猴子在潜意识中也会偏向种群内部的成员。它们会将水果等好的东西与内部成员联系起来，将蜘蛛等坏的东西与外部成员联系起来。

不幸的是，种族偏见不只是实验室里的研究现象。前文提到过，经济学家们发现写有典型白人名字（艾米莉、格雷格）的简历接到的招聘者电话比写有典型黑人名字（拉齐莎、贾马尔）的简历要多。更令人心寒的是，针对美国法庭笔录进行的研究显示，死刑案件中的受害人如果是白人，那么黑人被告比白人被告更可能受到死刑宣判，对于面部特征符合典型黑人形象的被告来说更是如此。种族问题在政治方面也是影响深远。经济学家赛斯·斯蒂芬斯–大卫德威茨（seth stephens-Davidowitz）统计了不同区域的人们在谷歌搜索中输入“黑鬼”（nigger）一词的频率，并将结果标记在美国地图上。结果显示，“黑鬼”一词在搜索中出现（通常是为了搜索关于种族的笑话）频率较高的地区，在2008年的总统大选中投给巴拉克·奥巴马的选票也很少，远远少于2004年约翰·克里在这一地区获得的选票。这样的种族嫌恶为奥巴马的对手送上了三到五个百分点的领先优势。在美国大选中，竞选人从自己家乡所在州获得的支持也不过如此，这个数据足以影响大多数总统选举的结果。

鉴于种族偏见的影响力之深，影响范围之广，有人可能认为我们已经对种族歧视“习以为常”。但细想一下，这个逻辑是说不通的。对于以渔猎为生的人类祖先来说，人们遇到所谓的“其他种族”的可能性非常小。事实上，住在山那边的“他们”在相貌和身体上也许与“我们”并没有什么差别。这就说明，种族问题并不是区分群体成员的本质标准，只不过在当今社会，我们恰好将其作为一个标准而已。从进化的角度来看，如果人类大脑中存在一个社会分类系统，我们希望这个系统能够更加灵活，能够根据语言、穿着等文化特征，而不是根据先天遗传获得的身体特征给人们分类。

由此，罗伯特·库尔茨班（robert Kurzban）与同事一起设计试验，研究人

们的种族意识和文化印记意识孰轻孰重。研究人员首先给参试者播放一段录像，录像记录了两支篮球队的队员互相争吵的场景，两支球队中都有来自不同种族的队员。随后，研究人员向参试者展示不同运动员的照片，配以偏袒己方的话语，如“是你们先开始生事的”。看完图片后，参试者在预先不知情的情况下被要求进行记忆测试，将队员的照片与他们所说的话匹配起来。通过分析参试者在记忆测试中做错的匹配，可以发现参试者对篮球队员的分类方式。种族意识强烈的参试者基本不会把白人说过的话和黑人说过的话弄混；同样，对球队身份十分敏感的参试者也不会把不同球队球员的话语弄混。库尔茨班和他的同事发现，如果试验材料中对于球队身份没有做出明显的标记，人们就会更多关注种族差异而不是球队身份。也就是说，人们基本不会把不同种族球员的话语弄混；来自不同球队球员的话语则相对难以区分。然而，如果球员穿上不同颜色的T恤衫，用以区分球队身份，所有的结果都会调换过来，种族差异一下子变得不那么重要，球队身份突然之间成了重点。

库尔茨班和同事们基于进化理论进一步提出假设：如前所述，因为种族差异并非根深蒂固的进化类别，这种基于种族差异而做出的分类是可以变化的。但同样的逻辑并不适用于性别差异（男女差异）。对于以渔猎为生的祖先来说，遇到男人和女人是再正常不过的事情，而且男人和女人的生理差异也十分明显。这说明与基于种族差异的分类相比，基于性别差异的分类更加难以改变。这也正是库尔茨班和同事们的研究发现：不论是黑人还是白人，不论队员身穿哪种T恤衫，参试者基本不会将男人说的话与女人说的话弄混。

这个试验说明，人类很容易根据群体成员的随机印记对他人分类，但这个结论本身并没有揭示这种分类对人类究竟有什么用处。亨利·泰弗尔（Henri Tajfel）与同伴的一个经典研究表明，社会分类会非常自然地成为社会偏好的基础。泰弗尔将参试者带进实验室，采用随机标准将所有人分为两组。在一次试验中，他首先组织参试者完成一项估计任务，然后假装将参试者分为“估计过高”和“估计过低”两组。（但事实上，分组是完全随机的。）随后，泰弗尔要求参试者以匿名的方式，为所有参与试验的成员分配钱财。他发现，尽管小组成员没有共同的经历，小组在未来不会继续存在，甚至连分组标准都是无关痛痒的，但人们在分配过程中依然倾向于对同组成员加以照顾。事实上，即使是公开地以随机方式分组，人们仍然会照顾组内成员。照顾组内成员并不仅仅是打破平衡的一种方式。人们通常会给组内成员多分钱，而不是给组外成员少分钱。

最近，人们开始将部族主义与特定的神经系统联系起来。如前文所述，催产素是一种神经递质，也是哺乳动物怀孕生产过程中十分重要的一种激素，会增加母体的同理心和对他人的信任。事实上，催产素（有时被称

为“抱抱化学物质”）会导致更加明显的社会偏好，甚至超过了我们所想象的程度。最近，卡斯滕·德·德勒（Carsten De Dreu）与同事们发现，向男性鼻腔内喷洒催产素能够促进组内合作，但无法促进组间合作，尤其当人们对组外成员心存恐惧时，效果则更加不明显。内隐联想测验显示，催产素还会使人们对组内成员的偏袒更加明显，对于组外成员的排斥感也有微小的增加。此外，催产素还会影响人们在道德两难处境中的选择，使人们更加不情愿牺牲组内成员利益，但对组外成员的感觉则不受影响。

总之，人类大脑中存在部族主义的思考机制。我们会自动把这个世界分为“我们”和“他们”，并且更加偏袒“我们”。长期以来，人类一直将语言线索作为辨别群体成员身份的可靠印记。从婴儿时期开始，我们便开始使用语言线索为周围人分类。现代社会中存在着种族歧视现象，但种族并不是根深蒂固的、先天的心理学分类，它不过是众多身份印记中的一种而已。正如泰弗尔的试验所揭示的那样，我们常常会基于极其随机的标准，将人们分为“我们”和“他们”。尽管这听上去有些疯狂，但很多情况下，这就是事实。对于一个需要大量成员共同合作才能够生存的物种来说，由于合作成员过多，如果没有文化身份印记的帮助，便无法相互辨别身份。如此来看，这种分类方式便合乎情理了。

在下一个话题开始前，还有一点需要说明：人类大脑虽然拥有部族主义的思考机制，但这并不意味着这种机制无法改变。增加阅历和主动学习都能够改变这种机制。更重要的是，大脑中存在很多不同的回路，有些回路的调整适应能力较强，有些则略差，它们之间也会相互竞争，争夺对人类行为的控制权。在后面的章节中，我们会进行更详细的介绍。

## 利益攸关方

亚特和巴德可以利用可信的威胁，诱导对方进入神奇角落。一个强大的第三方也可以达到同样的目的。例如，假设亚特和巴德同属于一个犯罪集团，犯罪集团内有一位首领，开出了一个禁止拒绝的条件：“如果你出卖同伙，我就把你杀死。”和以前的笑话一样，我把这种威胁称为“条件”，但其实真正的条件也能够达到同样的效果，比如：“保持沉默，我不会让你吃亏的。”

毫无疑问，牢固的合作是历史发展的动力之一：首领、国王和君主都利用自己手中不断变大的萝卜和棒子促进合作，促进生产（并总结出有效的手段）。17世纪英国的哲学家托马斯·霍布斯（Thomas Hobbes）认为，这是一件好事。他将国王赞为维护和平的利维坦（权威）和尘世间的上帝，使人类脱离自然状态下“肮脏、粗野、短暂”的生活。

利维坦不一定是尘世的上帝。在信徒眼中，一个超越自然的权威是维护合

作的最佳人选，因为超自然的存在可以是全知全能的，能够保证合作者受到最大限度的奖励，不合作者受到最大限度的惩罚。正如戴维·斯隆·威尔逊（David sloan Wilson）所认为的那样，宗教可能是文化演变过程中产生的一种机制，用于促进庞大群体内部的合作。敬畏上帝和做好合作者之间有着密切的联系，当然，这种观点存在已久。长久以来，信徒们对“不惧怕上帝”的人们总是心存戒备，这种心理也会一直延续下去。

从进化的角度来看，牢固的合作之所以能够达成，是因为它要求所有人按照简单的利己主义原则行事。群体中的成员进行合作，从而获得首领奖励，规避惩罚；而整个群体通过合作，也能够产生更大的收益，首领，特别是尘世中的首领，便可以从中获益。有人仍会质疑，如果没有权威的约束，从合作中获利的第三方是否真的能通过奖励和惩罚保证合作的稳定？这个问题十分重要。因为人类学研究显示，农耕社会以前，虽然没有权威为每个人分配任务，当时的社会状态依然是人人平等。

再回头看看“守口如瓶银行劫匪联盟”。按照我们之前的想象，联盟中的成员生来便都是守口如瓶的人，他们需要防止联盟外的劫匪利用自己的成员牟利。但事实上，联盟内部也可能存在不愿合作的个体。现在联盟面临的挑战就是：如何在没有强大首领的情况下，约束其内部成员遵守规则。联盟成员能够完成自我监督吗？

如果联盟规模不大，不合作的行为会立即招致以牙还牙的报复和惩罚：如果亚特出卖了巴德，巴德可以惩罚亚特，于是亚特下次就会选择合作。因为巴德通过惩罚亚特获得了直接利益，因此将这种模式称为直接互惠。但根据我们的假设，如果这个联盟十分庞大，那么巴德对亚特的惩罚便没有价值，因为两人很可能不会有机会再次合作。此外，不管怎样，如果联盟内的成员愿意在无法获得个人收益的情况下，对背叛者做出惩罚，便会极大地促进合作。如果联盟内部有足够多的人愿意充当惩罚者，那么背叛行为将会近乎绝迹；当然，惩罚行为也会变得难得一遇，因为这种情况下，因出卖别人而招致惩罚的可能性会非常大。

对不合作行为做出惩罚的整体意愿，是间接互惠的一种表现。之所以称为“间接”，是因为联盟成员在施罚时会付出直接代价，但收益却是间接的，是因联盟中其他成员实施惩罚而获得的。如果你由此想到了最初的公地悲剧，这是十分正常的。这种形式的间接互惠本身就是一种合作，将群体利益置于个人利益之上，是利他主义的一种形式。因此，这类惩罚通常被称为“利他惩罚”。这个名称可能会造成误解，因为利他惩罚者的行为不一定出于对群体利益的考虑。有的利他惩罚者只是单纯地享受严厉施罚的感觉，受罚者究竟是辜负了自己还是辜负了别人则意义不大。为了避免歧义，我将这种代价高昂的惩罚称为“亲社会”惩罚。

你是亲社会惩罚者吗？做个测试看看吧：假设在几千英里外有一个外国城市，那里发生了连环强暴谋杀案，罪犯已经伤害了十几名妇女和女孩儿，在被捕以前，他是不会收手的。如果匿名支付25美元就能确保将案犯绳之以法，你愿意吗？如果不是25美元，而是1美元呢？如果你的回答是肯定的，那么恭喜你，你是一名亲社会惩罚者。你愿意牺牲个人利益，换取他人的合作。（再次注意，不侵略也是一种合作。）

许多试验已经证明，人类的确属于亲社会惩罚者。其中最为著名的试验当属恩斯特·费尔（ernst Fehr）和西蒙·加士德（simon Gächter）合作的“公共财产游戏”。这是一个多人场景下的“囚徒困境”，与公地悲剧的原理相仿。研究人员先给每位参与者分发一笔钱，然后将他们分成不同的小组。每轮游戏中，每个参与者都可以选择向公共资金池放入一笔钱，进入资金池的总额会增加一倍，然后再平均分配给每位参与者。整个过程都是匿名的。

这种场景下，对整个小组来说，最合理的做法是让所有参与者把全部资金都放入资金池。这样可以使最多的资金获得加成，从而使参与者获得最大的资金回报。比如，假设参与者共有4位，每人的初始资金为10美元，那么4人的全部资金就是40美元。如果研究人员将池内资金增加一倍，变为80美元，那么每位参与者可以获得20美元，这一回报相当不菲。然而，从个人角度来看，最合理的（如果人是自私的）做法则是不投入任何资金，“不劳而获”地分享其他参与者的投资回报。不劳而获者能够保有自己最初的资产，同时获得公共资金池里的分成。如果4人中有1人不劳而获，那么一轮结束后，不劳而获者手中拥有25美元，而其他人的资产仅仅增加了5美元。公共财产游戏中，不劳而获的行为相当于囚徒困境中的背叛行为，或者是公地悲剧中随意增加牲畜数量的行为。

典型的公共财产游戏反复进行多次后，很多参与者开始合作，他们至少会将一部分资金放入公共资金池。但也有一些人开始不劳而获，他们向资金池中投入很少的钱，甚至分文不出。选择合作的参与者发现自己被利用，于是开始减少甚至停止投资。随着游戏轮数的增加，进入资金池的数额越来越小，更多的参与者会说“见鬼去吧！”。最终，每轮的总投资额几乎降到了零，真是悲剧。

但是，如果选择合作的人有机会对不劳而获者施以惩戒，那么情况便又不相同了。在这里，“惩戒”的意思是“有代价的惩戒”，也就是拿出一笔钱，使另一位参与者的收益降低。例如，一轮游戏后，一位参与者可以拿出1美元，将另一位不劳而获者的收益降低4美元，相当于在经济上给人以当头一棒。研究人员引入这个惩罚机制后，进入资金池的资金总额通常会增加。重要的是，所有人都清楚施罚者并不会从中受益，但即使这样，依然有人愿意承担施罚者的角色。此外，惩罚机制对合作行为的促进作用通常

是立竿见影的，甚至不必有人亲自实施惩罚，合作的状态便已经有所改善。这说明，即使施罚者并不能得到物质回报，有不劳而获想法的人们也会担心自己因此受罚。

人类为什么是亲社会惩罚者？人们为此问题争辩不休。有人认为亲社会惩罚不过是直接互惠和信誉管理在进化过程中的副产品。也就是说，大脑会自动将所有人认定为潜在的合作对象，并认为我们的行为总会受到别人关注，因此我们会惩罚与自己无法再次合作的人。对于以渔猎为生的小型团体来说，这种假设并非毫无根据。还有人认为，亲社会惩罚通过群体的生物选择或自然选择进化而成。也就是说，亲社会惩罚是对群体有益的行为，通过实施亲社会惩罚，个体可以帮助自己所属的群体战胜其他群体。此类争论十分有趣，但我们无须在此表明立场。重要的是，亲社会惩罚的行为真实存在，并且与我们所熟悉的一种心理学现象十分相符。

如你所想，情感就是亲社会惩罚的驱动力。费尔和加士德向游戏参与者询问，如果他们在实验室之外遇到了不劳而获者，会有怎样的感觉。大多数人表示自己会感到愤怒。如果自己是不劳而获者，他们认为别人也会对自己感到愤怒。我们将这种典型的道德层面的愤怒称为义愤。

人与人之间的相处方式是我们十分关心的话题，人类对于小说的迷恋便是对此最好的注解。如果人类完全自私，我们便不会甘愿支付不菲的价钱，只为听一个虚构的故事：一群社会底层的孤儿，拥有城市生存智慧，性格古灵精怪，与犯罪团伙斗智斗勇，最终取得胜利。我们之所以会沉浸在虚构的英雄与恶人的故事中，是因为这些故事触动了我们的社会情感。我们在实际生活中对待合作者和无赖的态度，也正是由社会情感决定的。在此，我们并非毫不相关的路人。

## 道德机制

不论是简单的细胞，还是像人类一样的超社会动物，地球上生命的故事其实就是越来越复杂的合作故事。合作造就了今天的我们，维系合作也是我们所面临的最大挑战。面对挑战，道德便是人类大脑给出的答案。［乔纳森·海特所著《公正的头脑》（*The Righteous Mind*）中，对这个问题进行了更加生动广泛的讨论。］

亚特和巴德的故事告诉我们，很多互补策略都能使性本自私的个体进入神奇角落，选择合作。在囚徒困境中确保合作成功的策略适用于一切社会困境。任何情况下，“我”和“我们”的利益发生冲突时，都可以使用这种策略。例如，亚特和巴德的合作策略可以帮我们避免公地悲剧，因为相互关心的牧民会自愿控制畜群数量。或者有一位利维坦式的权威牧民监督其他牧民，确保他们遵守规则。抑或是牧民们可以跟踪观察内部背叛者的表

现，防止外部剥削者的介入，从而维系合作等。更重要的是，这些策略有助于解决真实世界中的各种问题：对付贪婪牧民有效的方式，对付逃税者也同样适用。对付黑心商人、污染环境的违法者、侵略者以及伪装成朋友的敌人，都可以采用相同的方式。（下一章会对此详加解释。）

针对每一种合作策略，道德思维中都有一套情感机制，用以落实相应的策略。让我们来看一下：

关心他人：如果两位囚徒能在考虑自身利益的同时，略微考虑对方的利益，两人便能进入神奇角落。针对这种策略，人类发展出了同理心。概括来说，人类情感中包含关心他人的成分，特别是对于家人、朋友和爱人的关心。同时，这种情感机制还使我们不愿直接或有意地伤害他人，甚至（退一步讲）不愿令他人受到伤害。我将这种机制称为起码的尊重。

直接互惠：如果两位囚徒意识到，两人都能从未来的合作中获利，但倘若当下不合作，未来的合作也无法实现，两人便能进入神奇角落。针对这种策略，人类发展出了众所周知的愤怒和厌恶等负面反馈情感。这种情感促使我们对不合作者施以惩罚，或者对其避之若浼。同时，在一个难免犯错的世界中，宽恕应运而生，对负面的情感机制进行调解。对于积极合作者，人类也会通过感激，进行正面的情感激励。

可信的威胁与承诺：如果两位囚徒相互承诺对不合作者施以惩戒，两人便能进入神奇角落。这种策略所对应的情感是报复。众所周知，惩罚不愿合作者这种情感机制为多数人所拥有，即使这样做得不偿失，我们也在所不惜。同样的，如果两位囚徒下定决心惩罚自己的不合作行为，两人也能进入神奇角落。这种策略所对应的情感是荣誉感，由此产生的内疚和羞愧等自我惩罚情感也广为人知。与此相关的情感还包括忠诚，包括对爱情的忠诚。对更高权威的忠诚还包含了谦逊的美德和敬畏的能力。

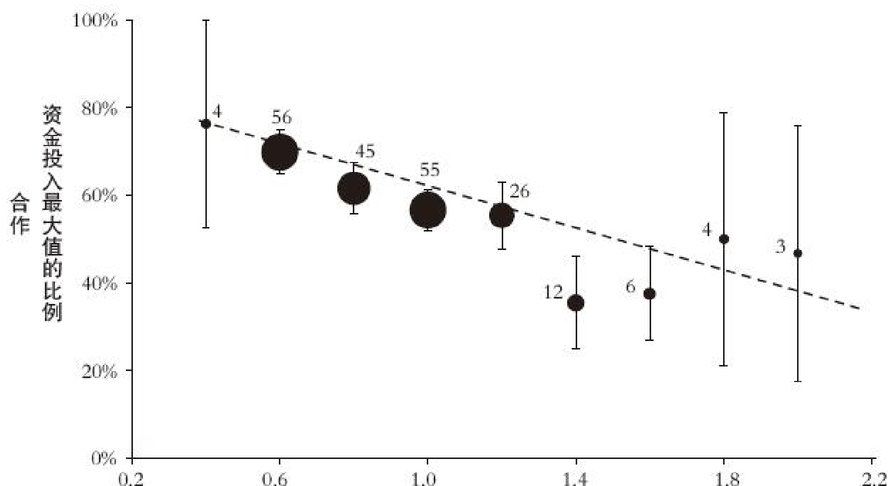
声誉：如果两名囚徒意识到，当下的不合作会导致其他知情人今后拒绝与自己合作，自己无法再从合作中获益，两人便能进入神奇角落。与这种策略相对应，人类从婴儿时期起，就有了论断是非的能力。我们会观察人们彼此如何相待，并随之调整自己的行为。此外，我们倾向于不负责任地制造并传播谣言，这种行为放大了论断之言的影响。因此，我们对来自他人的关注十分敏感，我们自己也变得十分自觉。如果自觉意识约束失败，我们在做坏事时被抓了现行，便会表现出明显的窘迫，向他人表明我们今后不会再次犯错。

归属感：如果两名囚徒同属于一个合作团体，并且团体内部成员能够准确辨认彼此的身份，两人便能进入神奇角落。针对这种策略，人类发展出了部族意识，对团体内部成员发出的信号高度敏感，潜意识里倾向于优待内部成员。即使是素不相识的成员，也能享有优于外部成员的待遇。

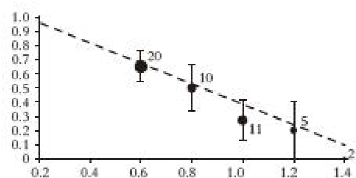
间接互惠：如果两名囚徒之外，还有另外的人对合作给予奖励，对不合作施以惩罚，两人也能进入神奇角落。与这种策略相对应，人类成为亲社会惩罚者，即使我们并不能从中受益，也会在义愤的驱使下对不合作者施以惩罚。同样，人们也期待他人对不合作的行为表现出义愤。

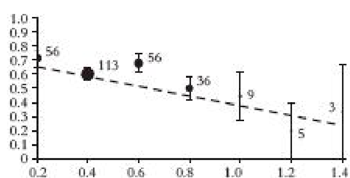
同理心、亲情、愤怒、厌恶、友谊、起码的尊重、感激、复仇、爱情、荣誉感、内疚、羞愧、忠诚、谦逊、敬畏、论断是非、说长道短、自觉、窘迫、部族主义以及义愤，这些都是人性中常见的特征*。社会中的每个人对这些特征的含义和作用都有自己的一套理解方式。但迄今为止，我们都没能弄懂这些迥然各异的人性心理特征何以能够相互契合，他们存在的意义究竟为何。上述所有心理机制的设计都近乎完美，能够促进本性自私的个体相互合作。这些策略的合理性可以通过抽象的数学语言证明，也可以通过被囚银行劫匪的例子印证。但不论是从生物学角度还是文化角度，目前我们都无法证明这些心理机制确实是为了促进合作进化而成。如果这些心理机制的进化并非以促进合作为初衷，那这种完美的设计便一定是不同寻常的巧合。

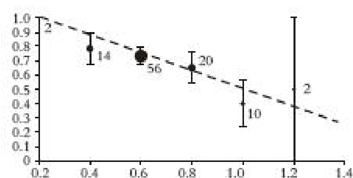
按照这样的道德观点，合作是发自本心的，人类无须逻辑论证便会选择合作。事实上，人类的情感已经替我们完成了逻辑的论证。为了证实这一观点，戴维·兰德（David rand）、马丁·诺瓦克（Martin Nowak）和我共同进行了一系列研究。首先，我们对已发表论文中以囚徒困境和公共财产游戏为基础的试验数据重新分析，对人们做出决定所需的时间进行特别关注。经过大量分析，我们发现了相同的规律：人们越快做出决定，就越可能选择合作。这一发现符合“合作是发自本心的”这一观点（参见图2.5）。




反应时间



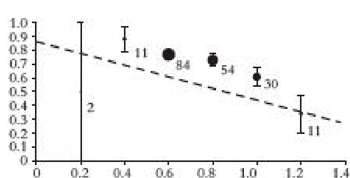




图2.55 组合作试验中的反应时间数据：迅速做出决定的人更有可能将集体利益置于个人利益之上。这说明，合作（至少在某种程度上）与自私相比，是更加感性的

随后，我们自己设计进行了公共财产游戏试验。在试验中，我们迫使一些人迅速做出决定（10秒钟以内），而规定另一些人必须缓慢决定（思考10秒钟以上）。不出我们所料，强迫人们迅速做出决定会使他们更易合作，而迫使人们进行缓慢的决定会使他们更加不合作（不劳而获行为的可能性增加）。在另外一组试验中，我们要求参与者在游戏前写下一件自己因直觉受益的事，或者举出一个自己被理性分析引入歧途的例子。回想感性思考的优势（或是理性思考的劣势）会使人们更倾向于合作。同样，回想理性思考的优势（或是感性思考的劣势）会使人们更加不愿合作。这些研究

再次印证了本章的重点：深植于人类道德思维的，是允许和促进合作的自动心理机制。

（注：通过这些试验，你也许会做出归纳：直觉是一切善行的来源，理性分析则是道德的敌人。但这种看法是错误的，事实上，我写这本书的目的便是要纠正这类错误。这些研究实际上表达出的信息是：社会直觉能够十分有效地避免公地悲剧。如前所述，公地悲剧并不是我们面对的唯一悲剧，稍后我们会对此详加阐释。）

我一直将本章中提到的心理机制称为“道德机制”。然而将“道德”等同于“促进合作的”仍然有些不妥。原因有二：首先，很多典型的道德现象看上去与合作并没有什么关系。例如，在某些文化中，食用某些食物或进行某些双方自愿的性行为是不道德的。这些禁忌与促进人们合作有什么关系呢？

需要澄清一点，我将人类思维中促进合作的心理工具称为“道德机制”，并不意味这种机制只能用于促进合作。事实上，我想表达的意思是，大脑之所以拥有这种机制，是因为它能够促进合作，但这并不代表该道德机制不能拥有其他用途。比如，鼻子可以起到架起眼镜的作用，但鼻子最初的进化并不是为了架起眼镜。同样，对同性恋的义愤也许不能促进合作，但人类产生义愤的能力依然存在，因为义愤对促进合作意义重大。也就是说，一些道德行为看似与合作无关，但实际上却与合作关系密切。例如，印度教中禁食牛肉，人们便不会将奶牛一次性杀掉吃肉，这会使奶牛成为长期奶源，保证食物供应。新教教义中要求人们高效工作、抑制消费，这就会使整个社会获得更多的资源。即使是对于自慰这种私密行为的禁忌也可能会起到一定的社会作用：对获得性满足感的其他渠道进行限制，教堂等合作机构的权利便得以扩张，因为教堂等机构是唯一能够祝福婚姻的地方。

其次，我提到的“道德机制”中，有些也许界限模糊，有些甚至非常不道德。关心他人无疑是道德的，无私地执行合作规则可能也是道德的。但直接互惠是道德的吗？对不合作者的躲避或惩罚也许会促进合作，但这种行为似乎并不十分道德。恰恰相反，这种躲避和惩罚似乎属于简单的利己主义。还有人类对复仇的渴望，虽然这种行为也可能会促进合作，但复仇远非我们所崇敬的道德行为。

事实上，我将这种心理机制称为“道德”，并不是想要表达我的认可，至少我对这种心理机制并非完全认可。相反，就像稍后会讲到的那样，我认为人类的大脑机制使我们陷入了许多不必要的麻烦。不管怎样，从描述性、纯科学的角度来看，尽管人类心理的很多方面并不那么美好，但这些不甚美好的方面都是人类心理机制的重要组成部分，是为了促进合作进化而成的。更重要的是，我们要知道，对所有确凿无疑的道德行为来说，这种心理机制才是它们在世间的起源。也就是说，并非所有为促进合作而产生的

行为都能获得“道德”的美誉，但如果人类大脑不曾为了合作而做出调整，那么世间一切被誉为“道德”的行为都不可能存在。

那么，人类大脑为何会为了合作而进行调整呢？也许这是上帝的安排，抑或只是大自然中的一个巧合。但不管怎样，我们不必再在神的意志和自己的运气之间艰难抉择。或许，人类之所以会产生合作思维，是因为合作能够带来更多的物质利益和生物资源，使我们的基因得以更多地传承下去。在进化的尘埃之上，一朵人类的善良之花徐徐绽放。

# 第3章 新草地上的争斗

新草地上牧民的思维中充满了促进合作的道德机制，但他们的生活却笼罩在部落间冲突的阴影之下。即使是和平时期，部落之间也分歧不断，在人应当如何生活这一问题上各执己见。这是为什么呢？上一章中，我们分析了能够促进合作的道德机制，以及能带我们进入神奇角落、避免公地悲剧的心理模式。本章中，我们将再次审视道德机制。这一次我们将探索为何道德机制会在现代社会频繁辜负我们的期望：为何人类的道德思维能够成功避免公地悲剧，却多次在常识道德悲剧面前无计可施？

## 冲突心理学

部落间合作的心理障碍主要来自两个方面：其一，在群体层面存在着一种古老而朴素的利己主义思想，或称为部族主义。它的表现形式是，人们总会优先考虑“我们”而不是“他们”。其二，部族主义之外，不同群体的价值观也大不相同，对“合作”一词的理解也有很大分歧。例如，南方牧民与北方牧民之间的分歧就已经超越了简单意义上的部族利己主义。崇尚个人主义的北方牧民坚信，勤劳智慧的牧民不应被强迫帮助那些愚昧懒惰的人；同样，推崇集体主义的南方人也坚持认为，让自己的部落成员在其他人丰衣足食的情况下忍饥挨饿是不对的，尤其是对那些遭遇不幸的人而言。即使不考虑利己主义，南北方牧民之间也存在着许多分歧。

部族冲突的两种表现形式自然地相互融合，不同部落会出于某些利己的原因，对某些价值观更加推崇。我把这种现象称为“有偏见的公平”。北方牧民是极端的个人主义者，南方牧民则是极端的集体主义者。那么更为中庸的东方人和西方人又是怎样的呢？假设东方牧场的土地比西方牧场肥沃，东方牧民也因此比西方牧民更加富有。由于东方牧民有能力向邻家兄弟伸出援手，他们也许会更倾向于个人主义，排斥集体主义；而西部则与之相反。两个部落的道德发展方向刚好相反，无论是东方人还是西方人都不会认为自己带有偏见。事实上，道德的转变过程可能需要几代人的时间，在社会应该如何构成这一问题上，某一个体的观点在有生之年并不会发生改变。

某些真正意义上的道德分歧从本质上来源于双方不同的着眼点。让勤劳智慧的牧民慷慨解囊，对愚昧懒惰的人施以无偿救济是不公平的，崇尚集体主义的南方人并非不明白这一点，他们对不劳而获的邻家懒汉也颇有怨言。只不过对南方人来说，任由部落成员在富庶年代死于贫穷太过残忍，即使是对于愚昧懒惰之人也是如此。同样，富庶的北方人对贫穷者也并非毫无恻隐之心，即使是好吃懒做的人，也经常能够以慈善捐助的形式获得

帮助。不过，北方人不愿被迫帮助愚昧懒惰的人，他们反对将接受帮助视为权利，将愚昧和懒惰变成合法行为。他们认为，这种行为会损害社会利益，甚至比任由部落成员死去还要糟糕。

群体间其他的道德分歧则与着眼点无关。某些群体内部的道德价值观有其自身特点，外人无法理解。在外人看来，这些价值观武断古怪；但在群体内部，这些做法却是理所当然、神圣不可侵犯。例如，某一部落禁止女性在公众场合裸露耳垂。但其他部落则认为女性裸露耳垂并无不妥，如果这条禁令让他们感到不适，他们也完全不会对其进行容忍。同样，有些部落为特殊的个人、机构、文本和神灵赋予道德权威和政治权威。比如，某个部落的圣书规定，白绵羊和黑绵羊不能圈养在一起。这一说法得到了部落最高领袖的首肯，他代表了众神之神，说的话便是绝对真理。这种分歧出于本质，与程度无关。因为其他部落无论如何都不会信奉这样的圣书、神灵和领袖。

由权威的个人、神灵及圣书引发的分歧会导致部落间对基本事实认识不一。某个部落的圣书中讲到，新草地原本是他们祖先的家园，其祖先在很久以前被迫离开。其他部落则认为这完全是自我杜撰，他们会问："证据在哪里？"得到的回答是："就在圣书中！"这类信仰是本土化的，与人们的虔诚密不可分，人们对以某些专有名词表示的个人、圣书与神灵深信不疑。用更为中性，大概也更为恰当的词语描述，这种信仰是区域性的。但信仰者们显然不这样认为，在他们看来，这些信念反映的是他们对普遍道德秩序的理解，而其他部落的成员出于种种原因无法理解这种秩序。

由此可见，新草地各部落间之所以争斗不断，是因为每个部落都自私地将"我们"的利益置于"他们"之上；也是因为不同部落看待世界的道德标准不同。接下来，我们将从心理学和社会学角度审视道德冲突。

## 部族主义

导致新草地冲突的最直接原因是部族主义，即（理所当然地）优先考虑本族成员的利益。本节的篇幅很短，因为毋庸置疑的是，人类拥有部族主义倾向，这一倾向会使冲突升级。关于部族主义倾向是否存在，人们并无异议，问题在于我们为何会拥有这一倾向。在我看来，有明显的证据表明，部族主义倾向是与生俱来的。人类学报告再次指出，对本族成员的偏爱和民族优越感是普遍存在的。幼童根据语言线索对本族成员进行辨认、产生偏爱。内隐联想测验揭示，大部分的成人、小孩，甚至猴子都会将外族人与负面概念关联起来。人们自然而然地对本族成员产生偏爱，即使以随意的标准划分临时小组，结果也是一样。人类联合心理学报告在进化报告部分预言，尽管联合分类体系能够轻易代替种族分类体系，但性别分类体系

却是无可取代的。在催产素这种神经递质的作用下，人们会有选择地偏爱本族人。总之，与外族人的合作是不断发展变化的过程，人们对此现象的所有解释都包含同一内容，即：与外族人相比，本族的合作伙伴更受欢迎。事实上，一些数学模型也表明，如果没有部落之间的敌对情绪，部落内部的利他主义也很难得以发展。

总之，部族主义倾向貌似是天性使然，但无论如何，这一倾向的存在都是不争的事实。当各个部落的人尝试共同生活时，部族主义倾向注定会引发各种矛盾——尽管这些矛盾并非“不可调和”。

## 合作的条件是什么？

部族主义使得群体间的相处变得困难，但部落内部的利己主义并不是唯一的问题。群体间合作的适当条件是什么？群体间是否应当彼此寄予期望？跨文化研究显示，不同群体对这些问题的看法截然不同。

在一系列意义重大的试验中，约瑟夫·亨里奇和他的同事与世界各地的人类学家组成了团队，这些人类学家来自非洲、南美洲、印度尼西亚、巴布亚新几内亚等地，专门从事小型社群研究。他们邀请小型社群的成员参与三项经济学博弈，希望衡量人们的合作意愿，检验他们对他人合作意愿的期望。上一章中已经介绍了前两项博弈：独裁者博弈和公共财产博弈，本章将介绍第三项：最后通牒博弈。

在最后通牒博弈中，某一参与者（提议人）需要与另一参与者（回应人）共分一笔钱财，对于提议人的报价方案，回应人既可以接受也可以拒绝。如果回应人接受提议人的报价方案（例如：我拿六，你拿四），那么就按照这个方案对钱财进行分配；如果回应人拒绝，两人便一分钱也拿不到。按照惯例，最后通牒博弈同样以匿名方式进行，其根本目的在于考量人们分配资源时的公平意识。提议人的出价高说明这个人愿意与他人分享资源，也许是因为提议人认为这样做是公平的，也许是因为提议人希望回应人与自己看法相同。出价低则说明提议人更注重个人权益，并希望他人能够尊重这种权益。对报价的拒绝则意味着：“你的出价太不公平，我宁愿牺牲个人利益也要表达我的想法。”

亨里奇和同事发现，最后通牒博弈中，不同社群的典型报价方案也各不相同。比如，一个较为极端的例子发生在秘鲁的马奇根加，他们向回应者提出的平均报价只有钱财总额的25%，而25名马奇根加回应者中，只有一人选择拒绝。这里的社群成员愿意拿出的资源很少，彼此之间的期望值也很低。马奇根加的情况与美国和其他西方工业国家的情况完全不同，这些区域的平均报价约为总额的44%，最普遍的出价比例为50%，低于20%的报价有一半会被拒绝。还有一些小型社群与西方社群大致相似。例如，津巴

布韦的一群村民平均出价比例为45%，约有一半的低报价遭到拒绝。与上述情况完全不同的是，巴拉圭的阿契人和印度尼西亚的拉梅拉若人平均出价比例都超过了50%，没有一例报价遭到拒绝。而在巴布亚新几内亚的au族，多数人的出价比例都超过了50%，但那里的回应人却往往会拒绝过于慷慨的报价，对于过低的报价也同样会拒绝。因此，对于不同地区的人来说，道德机制的作用方式也各不相同。

如前所述，公共财产博弈就是实验室版的公地悲剧。参与者可以向公共资金池中放入资金，研究人员将资金池中的资金翻倍，然后平均分配给每位参与者。参与者个人通过不投资（不劳而获）的方式可以获得最大收益，但整个群体却需要通过全额投资才能获得最大收益。在西方典型的公共财产博弈（将大学生作为参与者）中，平均贡献率约为40%~60%，大多数的参与者或者倾其所有，或者分文不出。（有趣的是，美国人在合作中对背景线索极其敏感。比如，试验中的囚徒困境被称为“华尔街博弈”还是“社区博弈”，会对合作结果产生巨大影响。）相比之下，马奇根加人在博弈中的平均贡献率只有22%，没有参与者将全部资金投入资金池。与西方人不同，在玻利维亚中部的提斯曼原住民中，极少有人分文不出，也极少有人倾其所有。在这里，我们看到了不同地域呈现出的巨大差异。

独裁者博弈并不是真正意义上的博弈，因为提议人能够完全掌控博弈结果。独裁者博弈中，人们在得到一笔钱的同时，也可以选择与他人分享这笔钱财，至于给多给少，给还是不给，由他们自己决定。在这种博弈中，西方大学生或者给出50%，或者选择不分享，这与他们在公共财产博弈中的表现基本相符。（这也说明，美国人的行为受背景信息影响严重。在独裁者博弈中，倘若还有第三个选择，允许人们从陌生人手中夺取钱财，美国人通常会拒绝分享。）同样，提斯曼人在两类博弈中也表现出了同样的文化印记，他们的平均出价比例为32%，一般不会分文不出。在肯尼亚的奥马和坦桑尼亚的哈扎，最为普遍的出价比例分别为50%和10%。我们可以想到，最具合作精神的社群也是最苛刻的社群，对于不合作者，他们严惩不贷。（独裁者博弈属于一厢情愿的博弈，并不牵涉合作问题，因此有人也许会怀疑，独裁者博弈与合作究竟有何关系？）

为什么不同文化背景的人在博弈中的表现大相径庭？正如我们所想，博弈方式反映的是人们的生活方式。亨里奇和同事以两种方式对这些社群的特点进行描绘。第一，他们对不同社群的“合作收益”进行排名，统计社群成员在合作中的获益程度。例如，马奇根加人以家庭为单位独立谋生，而印度尼西亚的拉梅拉若人则需要十几个人一起结伴捕鲸。由于经济生活的模式不同，在最后通牒博弈中，拉梅拉若人的出价是马奇根加人的两倍也就不足为奇了。试验人员还会根据“市场一体化”程度，对不同社群进行排名，也就是将人们在日常生活中对市场交换的依赖程度作为判断标准（例

如：获得食物的途径是购买还是自己生产）。第一章已经提到，参与市场经济是一种大规模的合作形式。亨里奇和同事发现，合作收益和市场一体化可以对不同文化间2/3的差异现象做出解释。近期的一项研究表明，市场一体化是衡量各个社群在独裁者博弈中利他主义表现的最佳指标。与此同时，参与者的性别、年龄、相对财富、可供分配的钱财等指标虽然看似十分重要，能够用于预测合作行为，但在实际的预测中却几乎没有用处。

在更加具体的层面，试验研究的结果与文化行为也十分契合。以巴布亚新几内亚的au和Gnau为例，这两个地区在最后通牒博弈中的报价比例普遍超过50%，人们通常还会拒绝过于慷慨的报价。我们发现，这些群体的文化中包含礼尚往来的传统。一个人接受礼物的同时，回赠礼物的义务也随之产生，这种赠礼行为还表明受礼者地位低于赠礼者。巴拉圭的阿契是在最后通牒博弈中表现最为慷慨的部落之一，几乎所有参与者的报价比例都超过了40%。阿契是极其崇尚集体主义的部落，成功的猎手们通常把捕获的猎物放在营地周围，告诉别人自己并没有捕获猎物，而其他人则会找到猎物，将其平均分给营地成员。肯尼亚的奥马也是推崇集体主义的部落，部落成员们自发地给公共财产博弈贴上了“齐心协力”的标签，意指大家为了修建学校和公路等共同的目标，共同努力。在公共财产博弈中，奥马人的贡献率高达58%。

近来，本尼迪克特·赫尔曼（benedikt Herrmann）和同事对一些大型社群中的合作和惩罚情况进行了调查，结果同样令人吃惊。他们邀请世界各地的城市居民反复进行公共财产博弈，并规定参与者有权惩罚不劳而获者。部分的调查结果如图3.1所示。

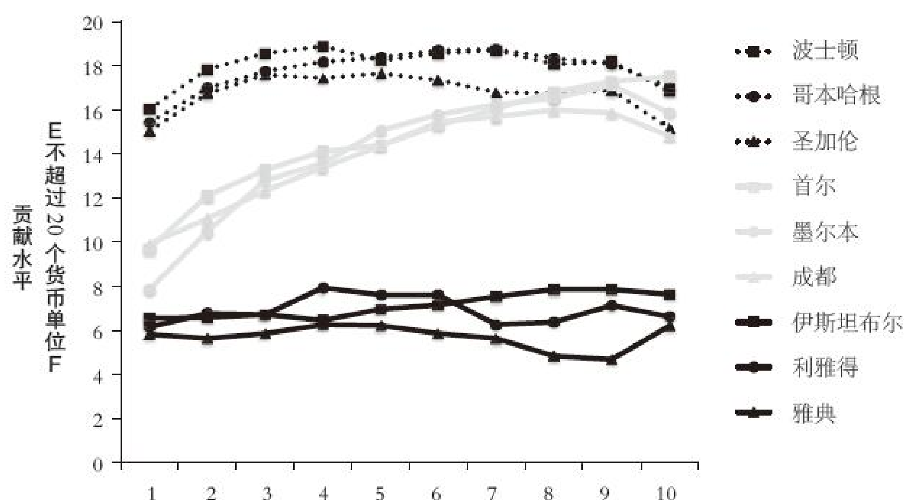


图3.1 不同城市的居民反复进行公共财产博弈，合作水平与合作曲线的轨迹差别很大

X轴表示博弈轮次（第一轮、第二轮……），Y轴表示平均贡献水平。首先，我们能够注意到，不同城市居民的贡献率起初的差异十分明显。雅典、利雅得和伊斯坦布尔居民的平均贡献率略高于25%，而波士顿、哥本哈根和圣加伦居民的平均贡献率却高达75%以上。其次，随着时间推移，博弈游戏呈现出三种不同的模式。其一，以哥本哈根为代表的城市在博弈中的贡献率始终较高。因为多数人从一开始便愿意合作，也愿意对少数的不合作者施以惩戒。（然而，这类城市中，如果人们没有机会实施惩罚，合作最终也同样无法延续。）其二，以首尔为代表的城市在博弈伊始的贡献率处于中等偏高的水平，但不劳而获者受到惩罚，不劳而获现象得到遏制后，贡献率便会大幅上升。其三，以雅典、利雅得和伊斯坦布尔为代表的城市在博弈中的贡献率始终很低。这种模式让人感到十分疑惑：这些城市的合作者明明可以对不劳而获者施以惩戒，这里的合作为何没能像首尔那样迅速发展呢？

我们发现，在雅典、利雅得和伊斯坦布尔等城市中存在一种反社会的力量。在公共财产博弈中，合作者可以惩罚不劳而获者，但不劳而获者也同样能够惩罚合作者，这种现象被称为“反社会惩罚”。在雅典等地，不向资金池中投资的人常常会出钱惩罚将资金放入公共资金池的人。人们为什么这样做？报复是一方面的原因，因为合作者对不劳而获者施以惩罚，后者便会怀恨在心，伺机反击。但报复并非唯一的原因，因为在有些地方，贡献率低的人甚至在博弈的第一轮就对合作者施以惩罚！他们好像在说：“见鬼去吧，你们这群慈善家！想让我按你们的规则办事，想都别想！”如图3.2所示，反社会惩罚的盛行成为群体合作失败的重要标志。

因此，在某些地区的公共财产博弈中，利他主义与亲社会惩罚等维系合作的力量反而被反社会惩罚压倒。这也说明人们的博弈方式似乎能够反映当地的文化特征。试验人员对每个城市数以千计的人们进行了“世界价值观调查”，并对其回答进行研究。结果显示，对偷税漏税和乘车逃票的行为持宽容态度的地区，反社会惩罚较为普遍。同样，我和同事在以公共财产博弈为模型的试验中发现，信任日常合作伙伴的人潜意识里也更愿意与他人合作。在我写作本书的过程中，由于希腊（图3.2的右下角）政府濒临破产，欧洲经济正深陷危机，不论是实验室数据还是现实生活都显示出同样的趋势。希腊的危机使整个欧盟也面临解体的风险，因此，丹麦（图3.2的左上角）等处于领导地位的国家便开始争论，为了更大范围的利益，是否应当对希腊进行经济援助，经济援助应当基于何种条件进行。

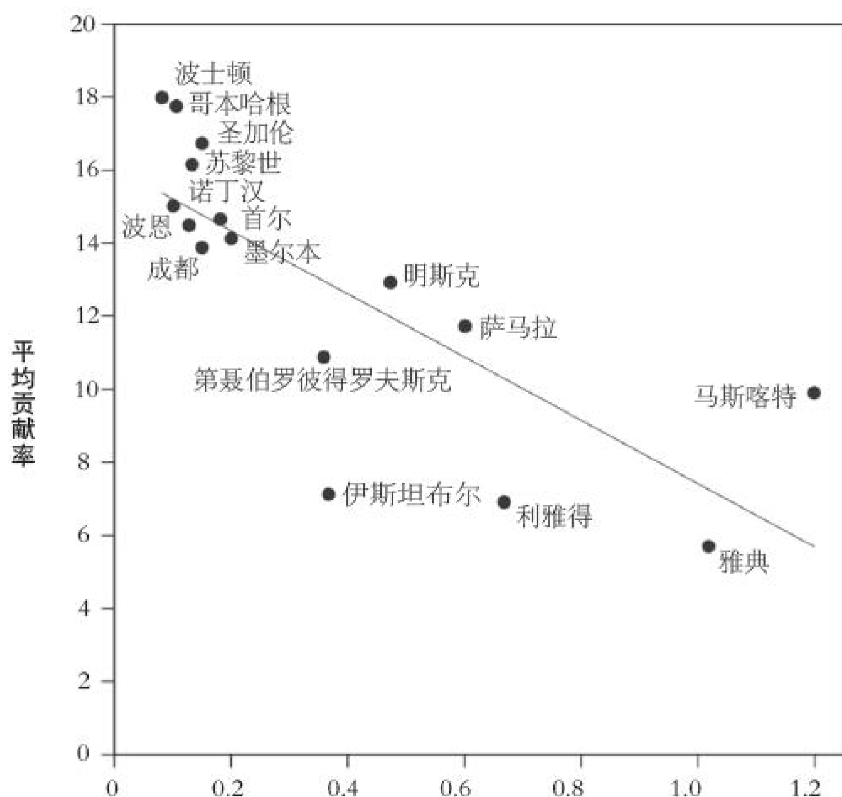



图3.2 试验人员在世界各大城市反复进行公共财产博弈，结果表明合作水平与惩罚合作者的“反社会惩罚”呈现负相关

（进一步展开讨论前，我想要声明，无论是本章还是其他章节，我都无意针对希腊人或任何国家和部落的成员。相反，我希望人类可以从各种社会制度的成败中汲取经验教训。单独的个体无须为社会制度的成败负责。然而，为了汲取教训，我们必须敢于建言，即使有些话可能会产生误解，令人感到冒犯；即使有些话听上去令人不快，像是存有偏见的观点，我们也在所不惜。）

## 尊严与和谐

20世纪90年代初期，德福·科恩（Dov Cohen）和理查德·奈斯比特（richard Nisbett）对美国内部的文化差异进行了一系列研究。他们邀请密歇根大学的多名男生进入实验室，进行“反应时间有限时，对某些人为判断的研究”。学生们逐个到达实验室，首先填写一些表格，然后穿过狭

长的走廊，将表格放到走廊末端的桌子上。按照试验人员的安排，走廊中间有一个人站在文件柜旁整理资料，每位学生交表格时都会经过他的身旁。学生们交完表格返回时会再次与他相遇，这一次他会“砰”的一声关上抽屉，用肩膀撞向学生，咒骂他们是“混蛋”。

对于这样的侮辱，来自不同地区学生的反应也各有不同。位于走廊的独立观察者们报告称，与美国北部的学生相比，来自南部的学生反应更为强烈，他们更不愿认为这是开玩笑，并更容易被此激怒。不仅如此，两组学生的心理反应也呈现出不同的特点。试验人员在试验前后分别对学生们进行唾液取样，化验结果显示，受辱南方学生唾液中皮质醇（与压力、焦虑和兴奋度相关的激素）含量的增幅大大高于受辱北方学生和未受辱的南方学生。受辱南方学生的睾酮素含量也呈现出较大幅度的增长。

试验后期，试验人员要求学生们阅读一个故事片段，并对此做出评价。故事是这样的：

> 舞会开始才20分钟，吉尔把史蒂夫拉到一边，似乎有些心烦。
>
> “怎么了？”史蒂夫问。
>
> “是拉里。我是说，他知道我们两个已经订婚了，可是他今晚已经两次向我调情了。”
>
> 说罢，吉尔返回人群，史蒂夫决定密切关注拉里。果然，不到5分钟，拉里就走了过去，想要亲吻吉尔。

试验人员要求学生们把这个故事的结局补充完整。被试验同盟（这里指试验人员秘密安排的同伴，而不是美国南部联盟的成员）侮辱的南方学生中，75%的学生补充的故事结尾包含了暴力或以暴力相威胁的行为，而未受辱的南方学生中只有20%给出了这样的故事结尾。相比之下，北方学生受辱与否和他们给出的故事结尾并没有直接关系。

科恩、奈斯比特及其同事还想探究侮辱是否会影响一个人在现实中的行为。为此，他们安排试验对象参与“懦夫游戏”。学生们受辱（或未受辱）后，会遇到另一名试验同盟，这名身高6英尺3英寸、体重250磅的壮汉正在快步穿过走廊。沿着走廊放有很多桌子，空间十分狭小，参试学生和身材魁梧的试验同盟无法同时通过，其中一人必须让路。高大的试验同盟穿过走廊时会与学生发生碰撞，而且仅在最后一刻才会让路。试验人员对学生们选择让路时与试验同盟之间的距离进行了测量。总体来看，受辱的南方学生选择让路时，与试验同盟之间的平均距离为37英寸，未受辱者让路

的平均距离则是108英寸。对北方学生来说，侮辱对他们的“胆怯点”没有产生任何影响。与此同时，未受侮辱时，南方学生会比北方学生表现得更有礼貌。未受辱的北方学生让路的平均距离为75英寸。

为什么南方与北方的学生面对侮辱（或未被侮辱）的反应差别如此之大？科恩和奈斯比特认为美国南部与世界上某些其他的地区一样，拥有浓重的“荣誉文化”，他们也据此对试验结果做出了正确的预测。与亨里奇等人一样，他们也从经济的角度入手分析：南方经济最初以畜牧业作为根基，许多早期住民都来自畜牧经济为主导的英国边缘地区。面对侵略者的巧取豪夺，由于其财产便于移动，牧民们显得尤其脆弱。（偷绵羊比偷玉米更容易。）如果没有可靠的法律保障，入侵的威胁就会增大，英国高地与美国南部某些地区的历史教训已经证明了这一点。牧民们必须坚守自己的家园，否则将会变得一无所有。此外，正如上一章所述，牧民们必须让外界知道他们守护家园的决心，不给侵略者想象的空间，因为一旦被视为弱者，即使花费大量的时间和精力守土自保，恐怕最终也难逃厄运。侮辱是对于尊严的挑战，一旦牧民们接受了侮辱，哪怕只是一点，都可能会被人看作示弱的表现。反之，脾气暴躁或以此形象示人的牧民在战略上反而更有优势。

在密歇根大学进行的这项试验中，来自南方的大学生虽然不是牧民，但孕育他们成长的文化中，传统的尊严观念十分重要。一直以来，南方的尊严文化对社会产生了深远影响。南方的谋杀率比北方高出许多，但这是因为由争端和冲突引发的谋杀事件在南方更多，并非人性恶毒使然。有研究表明，南方人看待暴力的态度与其他美国人并没有太大差异，但他们对守土保家和妻子受辱时的暴力有着更高的认可度。同样，对于在羞辱面前反应不够强烈的男性，他们也会对其嗤之以鼻。

南方的尊严文化似乎对美国的对外政策也产生了深远的影响。历史学家戴维·哈克特·费斯切尔（David Hackett Fischer）认为，美国南方一直都“有力地支持每一场美国对外战争，至于开战原因和交战对象则并不重要”。他把这种思维模式归结为“南方的尊严观和勇士道德观”。例如，英国与法国在1798年和1812年分别发生了两场战争，美国南方先是在1798年以极大的热情支持英国，而后又在1812年以同样的热情帮助法国。由于共和党和民主党轮流执政，地区性的政治倾向总会发生巨大改变，但美国南方对战争的支持却始终不变，并且超越了党派的界限。比如，反对罗斯福新政的南方民主党人士对罗斯福在第二次世界大战中的军事举措十分支持。同样的，哈里·s·杜鲁门和林登·约翰逊获得的南方支持也更多因为其反对苏联的对外政策，而并非国内政策。

尊严文化强调自力更生和人的自主性，在美国南方影响广泛，相较而言，东亚地区的集体主义文化则更重视相互依赖和群体和谐。奈斯比特和同事

认为，集体主义与美国南方的尊严文化一样，都是人们根据经济状况在文化上做出的适应，也就是基于合作农业的经济体制。基于这个想法，彭凯平（Kaiping Peng）、约翰·多丽丝（John Doris）、史蒂芬·斯蒂奇（stephen stich）和肖恩·尼科尔斯（shaun Nichols）向美国参试者和中国参试者分别提出了这个经典的道德困境：法官与暴民困境。这个困境的场景如下：

镇上发生了一起谋杀案，尚未确定哪位族群成员应当为此负责……由于镇上曾多次发生族群冲突和暴动，当地的警察局长和法官心里很清楚，如果他们不能立刻找到并惩处肇事者，镇上的人将会发起反族群暴动，使族群成员的人身财产蒙受巨大损失。警察局长和法官面临着两难的抉择：为了避免暴动，他们可以指控族群内一名叫作史密斯的无辜者，将其判罪并投入监狱；他们也可以选择继续追捕真正的杀人犯，任由暴动发生，之后尽其所能镇压暴动，直到罪犯落网……为了避免暴动，警察局长和法官最终决定错误地指控史密斯先生，将其判罪并投入监狱。他们的做法使得暴动并未发生，也避免了大规模的人员伤亡。

多数美国人认为，不论益处多么大，蓄意诬陷一位无辜者的想法都是骇人听闻的。众所周知，哲学家们总愿意听取争论各方的不同意见，但知名哲学家伊丽莎白·安斯库姆（elizabeth anscombe）却表示，她会与为法官辩护的人划清界限。“我不想与他争论；因为他的内心是腐烂的。”

鉴于这些话题十分敏感，在继续论述前，我想先澄清一些错误的想法。第一，上述试验与心理学试验一样，关注的都是群体平均值的差异。研究表明，一般来说，南方人比北方人更可能为了维护尊严而使用暴力，但这只是平均数据。南方人中不乏性格温和之人，北方也有暴躁易怒的性格，在此之间的所有性格也都能在两个文化群体内找到。第二，我并不想在这里称赞或谴责这些文化倾向，相反，我认为对于文化倾向的评价应当基于它们在环境中所起的作用，后文还会对此进行解释。如前所述，惩罚在合作的维持中扮演着至关重要的角色。由此可知，南方的尊严文化并非盲目崇尚暴力，而是对不合作行为的惩罚持有极其谨慎的态度。在一般情况下，南方文化也是强调礼貌和尊重的文化。我认为南方的尊严文化对美国的对外政策产生了非常重要的影响，但我并未妄言这种影响究竟是正面还是负面，因为我自己也不知道。美国南部对某些战争的支持也许对保持美国和其他国家的自由起到了重要作用。事实上，我在后文将要捍卫的道德哲学在某些人看来也许会过于倾向集体主义。

再次申明，我并非要对美国南方的尊严文化指手画脚，相反，我认为这些文化差异进一步诠释了道德多元化，反映出社会环境的多样性。两种文化都是合作文化，只不过合作的条件不同而已。中国的集体主义文化强调主动合作和必要的个人牺牲，而美国南方的尊严文化则注重被动合作（对他

人财产和特权的尊重），倡导人们对蛮不讲理的入侵和威胁予以迎头痛击。

## 有偏向性的公平

1995年，《美国新闻与世界报道》（*U. S. New & World Report*）杂志向读者提出了两个问题。一个问题是："如果某人起诉你，而你赢得了诉讼，他是否应当承担法律诉讼费？"85%的人给出了肯定的回答。另一个问题是："如果你起诉某人，却输掉了诉讼，你是否应当负担诉讼费用？"只有44%的人给出了肯定的回答。人们态度的转变说明，公平意识很容易受到私心的影响。这是偏向性的公平，而非单纯的偏见，因为人们的本心是向往公平的。假设这本杂志将两个问题同时提出，恐怕没人会说："如果我赢得诉讼，则由输家承担诉讼费；如果我输了，则有赢家承担诉讼费。"我们真心追求公平，但在多数分歧中，公平的选项可能有很多，我们便倾向于做出最适合自己的选择，实验室中的很多试验都记录了这种倾向。荷兰一篇论文的题目形象地归纳了试验结果的走势："绩效薪酬是公平的，当我的业绩更好时，这一系统也尤为公平。"

琳达·巴布科克（linda babcock）和乔治·罗维斯坦（George loewenstein）等人组织了一系列辩论试验，阐明了偏向性公平背后的心理机制。某些试验中，参试者两人一组进行辩论，讨论骑摩托的人被汽车撞倒后应获得怎样的赔偿。该虚拟场景以得克萨斯州某个已经审判完毕的真实案例为基础。试验开始时，研究人员随机指派同组的参试者分别扮演案件的原告和被告。法庭辩论开始前，两人要分别阅读长达27页的案件材料，其中包括证人的证词、地图、警方报告、真实案件中原被告双方的证词等。阅读完毕后，参试者需要猜想真实案件中法官判给被告的赔偿，这时同组两人所扮演的角色是已经确定的。如果猜测准确，参试者将获得金钱奖励。为了不影响随后的辩论，参试者的对手并不会得知其猜测结果。随后的辩论结束后，根据赔偿金额大小，参试者也会获得相应比例的现金奖励。赔偿金额越大，原告所得的金额也越高；而赔偿金额越小，被告所得的金额越高。赔偿金额从0美元到100000美元不等，每组两人辩论时间30分钟。谈判过程中随着时间的流逝，两名受试者都需支付"诉讼费用"，如果30分钟后双方仍未能够达成一致，则两人都需缴纳额外的罚款。

总体来看，猜测法官判定的赔偿金额时，原告的猜测平均比被告高出15000美元。双方的猜测差距越大，谈判就越是难以为继。换句话说，参试者的私欲扭曲了他们对现实世界的认知，更重要的是，这种扭曲会对后续的辩论产生重要影响。双方猜测数额差距较小时，只有3%的小组未能在规定时间内取得一致意见，而双方猜测数额差距较大时，高达30%的小组都无法在规定时间达成一致。另一个版本的试验中，辩论双方要先对法

官判定的赔偿金额进行猜测，之后才能得知自己在组中扮演的角色，这一次，规定时间内无法达成一致的比例从28%降到了6%。

这些试验显示，所有人都是有偏向性的辩手，但更重要的是，人们认为自己的偏向属于无意识的行为。原告认为法官的判定金额较高，而被告则认为金额较低，但双方并没有有意识地提高或降低赔偿数额。（再次提醒大家，金钱的刺激促使参试者尽可能做到猜测准确。）相反，在纠纷中所处的立场似乎会无意识地改变人们对公平的看法，也改变了处理信息的方式。在一个相关试验中，研究人员发现，人们对支持己方观点的案件材料印象更深。人们对公平的认知带有下意识的偏向性，即使是理智的人也很难达成一致观点，通常会导致两败俱伤。

为了探索人们在现实世界中对偏向性的公平的看法，研究人员查看了宾夕法尼亚州公立学校关于教师薪酬的谈判记录。这些谈判中，教师工会与校董事会的提议都使用其他有可比性的学区的薪酬标准作为参考。然而，究竟哪些学区是“有可比性”的，这个问题并没有确切的标准。研究人员猜测，关于教师薪酬的谈判可能会陷入更大的僵局，因为双方对有可比性学区的选择带有偏向性。研究人员分别走访了校董事会主席和教师工会主席，请他们列出附近有可比性的学区。与预期结果一样，与校董事会主席相比，教师工会主席列出的有可比性的学区中教师的平均工资水平要高出许多。随后，研究人员查阅了学区档案，发现如果教师工会主席和校董事会主席对有可比性学区的界定差异过大，该学区发生教师罢工的概率会比其他学区高出50%。

哈丁最初的公地悲剧中，所有牧民都处于均势。因此合理的解决方案似乎只有一个：把公共地平均分配给每位牧民。但是在现实世界中，各个利益集团不可能处于绝对的均势。事实上，即使哈丁已经对各场景进行了程式化的描述，如果不提出一些尖锐的问题，我们也很难对该场景拥有清晰的把握，比如每个家庭得到的牲畜数量相同吗？还是牲畜数量应当随家庭规模而变化？如此等等。只要人们站在不同的起点之上，便都会受到诱惑，有意无意地曲解公平的含义，使其更加符合自身利益。

金伯利·韦德·本佐尼（Kimberly Wade-benzoni）、安·坦伯伦塞（ann Tenbrunsel）和马科斯·巴泽曼（Max bazerman）组织进行了一项试验，从共同环境问题的角度阐释了有偏向性的公平。参试者在试验中扮演美国东北部海滨鱼群的股东，就过度捕鱼导致的日益严重的经济和环境问题展开辩论。在对照组中，辩论者各自代表不同的公司，但就像初始公共地的牧民一样，他们所拥有的财富大体相同。在关键的试验组中，辩论者对鱼群的收益结构各不相同，例如，尽管所有人都能从可持续的政策中收益，但一些人拥有相对长期的鱼群股份，另一些人拥有的股份期限则相对较短。辩论者的经济地位对等的控制组中，64%的辩论组就可持续发展方案

达成共识。当辩论者的经济地位不对等时，只有10%的辩论组能够达成共识。由此可见，相互冲突的个人利益大体对等时，人们很容易将个人利益放在一边，找到双方认可的解决方案。但当个人利益以不同的形式出现时，人们对公平的概念便会出现不同的理解，也更难达成共识。

具有讽刺意味的是，人们脑中有偏向性的公平概念根深蒂固，某些情况下，如果所有人都从私利出发，放弃高尚的选择，我们甚至可能过得更好。阿姆斯特丹大学的菲克·哈林克（Fieke Harinck）和同事们模拟了4个来源于真实刑事案件的案例，他们将素不相识的陌生人结成对子，就案例中的判决处罚展开谈判。每个小组需要同时谈判全部4个案件。在小组中，研究人员随机指定一人扮演辩护律师，为被告辩护，希望被告得到从轻处罚；另一人扮演检察官，希望能对被告严惩不贷。

每个案例中，对被告的处罚都有5类，从小额罚款到长期监禁轻重不同。每位参试者都会得到一份机密文件，站在己方立场（辩护律师或检察官）的角度，对每种处罚结果进行评估。研究人员将其中两个案件的结果安排为“零和博弈”，即一位参试者的获益必然伴随着另一位参试者等值的损失。对另外两个案件来说，双方则有可能找到“双赢”的解决方案，尽管一方的获益仍会导致另一方的损失，但这两个案例对双方的重要程度并不相同。参与者可以在对自己不太重要的案例中做出让步，而在对自己比较重要的案例中使对方让步。也就是说，这部分试验的设计能使双方同时获益，但前提是双方都愿意做出让步。在谈判者不知情的情况下，每种结果都对应一个预设的分值，能够反映该结果对谈判者是好是坏。研究人员将每组谈判双方所获得的分值相加，便可以衡量该组在寻找“双赢”方案过程中的表现。

上述做法是在为谈判试验制定一个标准。这项试验的奥义在于研究人员要求谈判者采用的谈判思维策略。有些小组被要求以纯粹自私的方式对待谈判，努力从轻或从重进行判处，因为该做法能够推动其职业发展、获得升职。另一些小组则被要求以道德的方式对待谈判：扮演辩护律师的参试者要寻求从轻判罚，因为轻判在这些案件中更为公正；扮演检察官的参试者则要寻求从重惩处，因为重判更为公正。

谁的最终表现更好呢，是自私的野心家还是寻求公正者？令人惊讶的是，自私的野心家表现更好。需要记住的是，自私的野心家并没有通过侵犯寻求公正者而获得成功，他们的成功确实是相互谈判的结果。哈林克和同事们发现，总体上说，与努力寻求公正的人们相比，以自私的方式进行谈判的人们更善于找到双赢的方案。这是为什么呢？

再次重申，在这一系列的实验中，谈判双方实现共赢的关键是在对自己不那么重要的问题上做出让步。自私的谈判者很愿意做出这样的让步，因为

让步会给双方带来净收益。你很清楚，你的对手也是自私的，只要能获得净收益，她也可以做出让步。这样一来，发现两人地位平等的事实后，两名自私而理智的谈判者便会愿意做出必要的让步，以扩大共同利益，然后再将收益平分。然而，如果谈判者的目的是寻求公正，而不是寻找对方的底线，那么很多模糊不清的考虑便会出现，有偏向性的公平也有可能随之产生。你的辩护对象也许确实应受轻判，你所判处的被告也许确实应受重罚，但这些案例中包含了很多合理的公平观点，你可以从中任选一个满足自身利益。相比之下，如果双方都只是简单地寻求对自己最佳的解决方案，那么剩余的回旋余地并没有太多，有偏向性的公平也不太可能使谈判陷入僵局。如果你把谈判过程视为双方共同寻求各自私利的过程，便不太可能认为谈判双方地位不对等，两位自私的谈判者不会抱有幻想，其自私是无处可藏的。这一令人吃惊的结果并不是要鼓励我们摒弃所有道德思维，追求纯粹意义上的自私，而是要提醒我们注意道德思维中存在的某些风险。在某些情况下，有偏向性的公平有害无益，我们最好能够先把道德放在一边，单纯地寻找最佳方案。

某些情况下，我们自己也许并不知道什么是公平，但我们采用部落中德高望重之人的观点时，我们的评判也可能产生偏向性。特别是制定公共政策时，这种情况更加有可能发生，因为普通市民几乎不可能获得足够的信息，有理有据地做出决策。杰弗里·科恩（Geoffrey Cohen）的一项试验阐释了有偏向性公平的部族主义特点。他找到的参试者中，有些人自认为是保守派，有些人自认为是自由派。科恩向他们展示了两份不同的社会保障政策议案：一个议案提供的福利待遇比现存任何社保政策都要优厚；另一个议案提供的福利待遇则比现存任何政策都要苛刻。如你所想，自由派比保守派更偏爱待遇优厚的议案，反之亦然。接下来的试验中，科恩找来一组新的参试者，依然包括自由派和保守派，然后向他们展示相同的议案。不同的是，这一次科恩声称两份议案分别来自民主党和共和党。如你所料，民主党支持的议案对自由派更有吸引力，而共和党支持的议案则对保守派更有吸引力。更令人感到吃惊的是，党派偏见的影响力如此之大，这个试验中，党派支持的影响完全抹杀了议案内容本身的价值。与包有保守派外衣的极端自由政策相比，自由派更加偏爱那些包有自由派外衣的极端保守政策。保守派也是如此，对他们来说，保守党派的支持比保守政策更为重要。如你所料，多数参试者拒绝承认自己的判断受到了党派偏见的影响。因为所有的影响都是潜意识层面的。

## 有偏向性的观念

对公平的判断在很大程度上有赖于我们对相关事实的理解，人们对2003年美国入侵伊拉克的态度就是个很好的例子。许多反对这场侵略的人无法理解支持的态度，他们问：“为什么是伊拉克？袭击我们的是奥萨玛·本·拉

登，不是萨达姆·侯赛因！”这些人不知道，或者不能完全理解的是，大多数美国人在那时都相信萨达姆·侯赛因本人也参与了恐怖袭击。同样，世界各地的许多国家对“9·11”事件也存在着不同程度的误解。2008年的全球民意调查显示，约旦、埃及和巴勒斯坦等国家的大部分人都相信，“9·11”恐怖袭击事件的背后推手并不是“基地”组织，而是另有其人（特指美国或以色列政府）。

为什么直接获取事实真相如此困难？一个原因便是由于单纯的自我意识导致的偏见。当事实扑朔迷离时，人们会选择相信符合自身利益的事实。一项著名的社会心理学试验中，来自两所不同大学的学生观看了两校之间足球比赛的片段。比赛中，裁判员做出了一些有争议的判罚，学生们则需要对这些判罚的准确性做出评判。如你所料，裁判员的判罚对己方不利时，该学校的学生就会把矛头更多地指向裁判。另一项经典研究中，试验所选取的参试者对待死刑持有鲜明的态度，试验人员向他们提供了正反两方面的材料，针对死刑震慑犯罪是否有效的问题提供了证据。我们也许认为，鉴于材料中包含了正反两方面的证据，参试者的态度也许能够变得更加温和。但事实似乎恰恰相反，参试者认为支持自己观点的材料比反方的证据更有道理。因此，权衡了关于死刑的正反两方面材料之后，支持者和反对者的态度都变得更加坚定。后来进行的另一项研究中，研究人员组织阿拉伯人和以色列人观看1982年贝鲁特大屠杀的新闻片段。尽管双方观看的是同一报道，但他们都认为这段报道对另外一方存有偏袒，研究人员将这一现象称为“敌意媒体效果”。再后来，丹·卡汉（Dan Kahan）和同事进行了一项研究，他们向参试者播放一段示威游行的录像，要求参试者判断，示威者究竟是在行使自己言论自由的权利，还是已经越过了法律界限，在进行阻挡或威胁行人等违法行为。其中一部分参试者被告知，示威者正在堕胎诊所门前抗议堕胎；另一部分则被告知，示威者在校园征兵地点外抗议军方对同性恋的“不问不说”政策。不管是反对堕胎还是维护同性恋的权利，对示威者所拥护的立场不感兴趣的参试者更倾向于断定示威者越过了法律界限。

这些场景中，人们利用模棱两可的描述，构建出符合自身利益的价值观。但有些时候，基于偏见构建的价值观却会与我们的自身利益相悖，或是看上去相悖。以人们对气候变化的态度为例，多数专家一致认为，人类活动是导致全球气候不断变化的主要原因，我们应当采取有效措施，让这种趋势减缓甚至停止。但许多人，尤其是美国人，都对气候变化的真实性持怀疑态度。这个场景中，人们无法获得事实真相并不仅仅是因为心怀偏见。某些个体通过否认气候变化的真实性，也许能够从中渔利，比如高碳排量公司的首席执行官等，但包括气候变化的怀疑者在内的多数人并不能从中得到好处。不管怎样，2010年美国盖洛普民意调查显示，只有31%的共和党人相信全球气候变暖正在发生，而66%的人则认为这一问题的严重性在

新闻报道中被夸大了。你也许会认为，针对气候变化问题，民主党人和共和党人都有理由努力控制气候变化，因为在地球未来的宜居性问题上，两个党派的利益完全没有高下之分。（他们也许还要拯救那些寄希望于在不久的将来进入极乐之境的人们。）既然如此，为什么有那么多的保守派完全忽视自身利益，否认气候变化的事实呢？从意识形态层面也许可以找到一个原因：总体上说，保守派对于通过集体努力解决集体问题的必要性持怀疑态度。这是问题的关键之一，但无法解释为何保守派对全球气候变暖的关注变弱，甚至不如刚刚过去的几年（稍后我会解释），也不能解释与其他国家的保守派相比，美国的保守派为何更加不关注全球气候变暖问题。

卡汉和同事认为，气候变化问题的关键在于认识到，获得事实真相本身就是公共地问题的另一种表现形式，这个问题本身既包括了个人利益，又包括了集体利益。直面气候变化问题并采取相应措施无疑符合集体利益，但对我们某些人来说，个人的支付矩阵会更加复杂。假设你所居住的社区中，人们不仅对气候变化持怀疑态度，还会对与他们意见向左的人提出质疑。这种情况下，你是相信气候变化好，还是做怀疑者好？作为一名普通市民，你个人的态度不太可能影响地球的气候，但却很有可能影响你与周围人的融洽相处。如果你本人相信气候变化问题，但生活在你周围的人都是气候变化的怀疑者，争议出现时，你有三个选择：一是心存疑虑，保持沉默；二是违心表达，隐藏自己的观点；三是冒着被排斥的危险，说出真实想法。如果你在赫伯的烧烤店里保持沉默，也许会有一丝机会改变人类历史的发展轨迹，相比之下，上述三种选择没有哪一种具有特别的吸引力，需要付出的代价也是显而易见的。卡汉认为，很多人之所以对气候变化持怀疑态度，并不是出于对地球自然环境的考虑，而是出于对自身所处社会环境的考虑，这样看来，他们的态度其实是非常合理的。这是个体理性对集体理性的一次胜利，当然也是潜意识中的胜利。

卡汉对这一问题的分析颇有预见性。传统观念认为，质疑气候变化的普通民众是愚昧的，通常不擅长批判性思考。根据这一观点，具有更多科学知识的人（科学素养高）和善于处理大量信息的人（计算能力强）则更有可能相信气候变化及其相关的风险。与此相反，卡汉的预测认为，与人们的科学素养和计算能力相比，文化观念（对本部落的忠诚）会对人们对于气候变化的态度产生更大影响。与传统理论不同，卡汉还认为，科学素养更高的人不会偏向真理，不论他们所属部落的观点如何，科学素养只会增强他们捍卫本部落观点的能力。

为了验证这些假设，卡汉和同事们选取了一大批有代表性的美国成年人，组织他们参与了一系列数学及自然科学测试。参试者还填写了调查问卷，旨在从两个维度考察参试者的文化世界观：等级制度与平权主义的维度和

个人主义与社群主义的维度。崇尚等级制度的个人主义者习惯推选位高者进行社会决策，同时对有损权威的集体行为十分警惕。相反，崇尚平权主义的群体主义者偏爱相对松散的社会组织形式，并对保护普通人权益的集体行为表示支持。在本项试验中，最重要的一点是：崇尚等级制度的个人主义者倾向于对气候变化持怀疑态度，而崇尚平权主义的集体主义者则倾向于相信气候变化是摆在人们面前的严重威胁，需要群策群力才能解决。

最后，研究人员向参试者询问他们对气候变化问题的看法。与传统的自由观念相反，研究者们发现，随着科学素养和计算能力的提高，人们对气候变化风险的感知会有略微下降。然而，把参试者按照观点分组之后，真相逐步显现了出来。不出所料，崇尚平权主义的集体主义者能够意识到气候变化带来的巨大风险，但在群体内部，科学素养或计算能力与感知风险没有呈现任何关联。同样，崇尚等级制度的个人主义者对气候变化带来的风险持怀疑态度，但在其群体内部，科学素养和计算能力出众的人反而对气候变化的风险存有更多质疑。（总体来说，科学素养和计算能力越高的人，对气候变化的质疑也越多，原因便在于此。因为崇尚等级制度的个人主义群体对整体结果产生了影响。）总体来看，科学素养或计算能力并不能很好地反映人们对气候变化所致风险持何种态度。相反，能够准确反映人们观点的是他们的整体文化观，即部落身份。（参见图3.3）。

需要澄清的是，你不应据此试验结果认为：人们对待气候变化的态度仅仅取决于身边朋友的态度，因此我们大可不必担心气候变化带来的威胁。（如果你愿意相信试验结果，便极有可能得出这样的结论！）比如，简的科学素养高于美国成年人的平均科学素养，但要想知道如何治疗牛皮癣，你并不会向她求助，而是会向专家求助，去咨询皮肤科医生。此项试验中的参与者并非气象科学专家，他们只是普通的美国成年人，他们的科学素养和计算能力呈正态分布，钟形曲线的中间数值代表了美国成年人的平均水平。尽管非专家人群对气候变化的态度存在普遍分歧，但多数专家一致认为，气候变化是真实存在的、严重的威胁。从试验中得出的结论并不意味着一切都是相对的，我们无法穿过文化的喧嚣，找到气候变化的真相；也不意味着普通人的思想只能受制于部落偏见。相反，在多数问题上，所有部落的成员都很乐意听取专家意见。（对部族的忠诚不能反映人们对治疗牛皮癣的看法。）从试验中得出的结论应当是：错误的思想一旦在部族文化中扎根，一旦成为部落荣誉的标志，就很难再改变，即使改变，也不是单靠教育能够解决的问题了。

“你认为气候变化对人类健康、安全和繁荣会产生多大的威胁？”

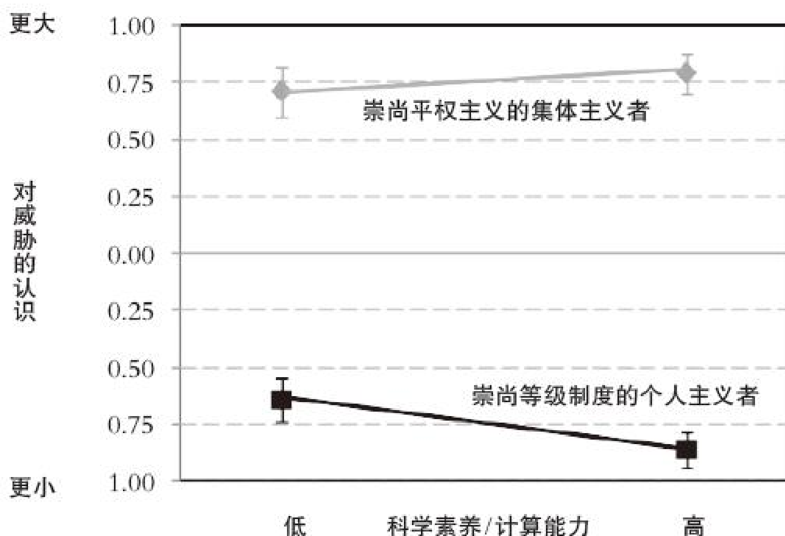



图3.3 科学素养和计算能力与普通人对气候变化威胁的态度几乎无关。人们倾向于与各自部落的观点保持一致

1998年，共和党和民主党同样相信气候变化问题已经出现。从那以后，证明气候变化的科学依据不断加强，但共和党与民主党对这一问题的看法却出现了重大分歧。到2010年，民主党中认可气候变化的人数达到了共和党人数的两倍。导致这种情况的原因并非共和党人的科学素养和计算能力发生了大幅下降，也不是因为民主党人吸纳了更多的科学精英。两党观点发生分歧的原因是，气候变化问题已经上升成为政治问题，迫使一些人不得不在相信专家和忠于部落之间做出选择。

需要注意的是，在其他问题上，与专家意见不一致的一方也许会是自由派。例如，对于深层地质隔离能否安全处理核废料这一问题，卡汉和同事发现，自由派（崇尚平权主义的集体主义者）更有可能否定专家的意见。在文化偏见方面，没有任何部落能够垄断。

## 有偏向性的升级

对本部落的忠诚可能会让我们对事实产生不同看法，其他的偏见也可能会影响我们感知世界的方式。伦敦大学学院的萨克文德·舍吉尔（sukhwinder shergill）和同事进行了一项简单的试验，他们针对偏见在冲突升级中的作用提出了一项假设，该试验便是为了验证这项假设而设计的。参试者需要成对进入实验室，第一位参试者的手指连接着一台小型按压机器，机器对

手指施加0.25牛顿的轻微压力，然后第一个人根据指令，以完全相同的力按压第二个人的手指。关键在于，第二个人并不知道指令内容。两人手指之间放有一个力量传感器，用来衡量按压的力度。随后，两位参试者交换位置，换由第二个人按照自己手指刚才受到按压的力度大小，以同样的力度按压第一个人的手指。然后两个人不断交换位置，互相按压手指，每次按压都基于自己在上一轮中受到按压的力度。在所有的试验组中，两人互相按压的力量都会迅速上升，最终几乎达到初始按压力量的20倍。（参见图3.4）

为何会发生这样的情况？说来奇怪，试验中力度的持续增加似乎与我们无法取悦自己这一事实相关。当你做出某个行为时，大脑会自动预测该行为的感觉结果，然后用这种信息抑制该行为的感觉效果。因此，自身主动行为引起的感觉比由他人引起的感觉更加不明显。（我知道你在想什么，是的，你想的是对的。）因此，当你做出按压动作时，手指所感受到的期望力量比他人施加给你的没有预期的力量要轻。也就是说，如果我们打自己一下，因为我们知道打击将会来临，所以真正感受到的力量会减轻。但如果我们被别人打了一下，我们的大脑无法获得来自于内部的提前预警，因此我们的感受也更加强烈。

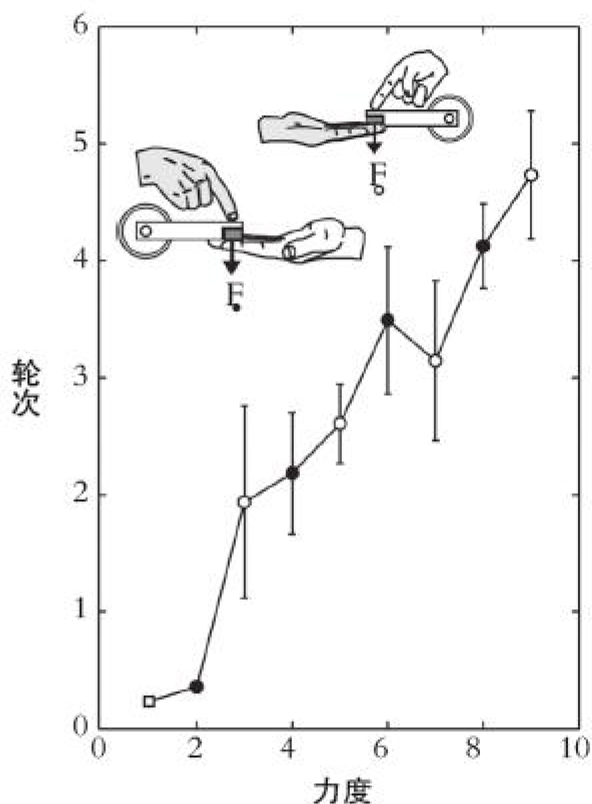



图3.4 每换位一次，尽管参试者试图维持之前的力度不变，但两人相互按压的力量都会增加

这项试验究竟是对真实世界中暴力行为不断升级所给出的解释，还是单纯的一个比喻，这个问题尚无定论。但其背后的机制却无疑是类似的：与我们按压别人时带来的痛苦相比，我们对自己被别人按压的痛苦更为敏感，如果这一情况不是无可避免，至少也是说得通的。媒体可被视为社会的神经系统，通过口口相传的方式传递信息，其宣传的内容大多是本部落的痛苦体验，而不是其余部落的痛苦体验。因此在某些情况下，道德偏见会成为世界认知体系的一部分。

这项认知原则解释了我们为何会低估自己造成的伤害，也解释了我们为何会高估自己带来的好处。尤金·卡鲁索（eugene Caruso）和同事找到了一篇期刊上的文章，分别要求4位作者对所有人在文章中所做的贡献进行评估。4位作者认为自己的贡献所占比例之和往往会超过140%。我们完全清楚自己所做出的贡献，因为我们亲自付出，但对于他人的贡献，我们却往

往知之甚少。

## 新草地上的生活与争斗

与新草地上虚构的部落一样，尽管价值观、信仰和利益并不相同，但我们依然正在努力，尝试在现代社会中共同生活。从历史的角度来看，现代生活总体很好。史蒂芬·平克在《人性中的善良天使》（*The Better Angel of Our Natures*）一书中提到，过去的几十年、几个世纪，甚至几千年中，人类的暴力行为大幅减少。他认为这种趋势之所以会出现，是因为文化使得我们思考、感受及组织社会的方式产生了深刻变革。变化主要包括：对民主治理的选择、对武力的使用进行合法垄断、培养同理心的娱乐活动、弱势群体的合法权利、作为可证实知识来源的科学、互惠互利的贸易等。前文提到的很多令人惊讶的研究结果也反映了这一趋势。需要再次强调的是，生活在市场一体化社会中的人们并非贪婪到无药可救，他们倾向于更加无私地对待陌生人，变得更加善于与人合作。

因此从人类历史的空想观点来看，新草地生活中出现的问题，有90%已经解决。但不管怎样，从地面向上看的视角十分不同，这种视角突出强调了尚未解决的10%的问题。尽管我们取得了长足的进步，但人们对此进步的理解依然不到位，我们所面临的问题也十分严峻。现代社会悲剧引起的、能够避免的苦难一点不比过去年代的苦难少，而且这些苦难的范围甚广，我们有限的思维无法真正理解。

本书的引言部分，我强调了我们所面临的一些问题：

> 贫穷：超过10亿人生活在极端贫困之中，为生存而挣扎。和贫穷有关的问题包括饥饿、营养不良、缺乏饮用水源、卫生条件差、暴露在有毒污染物之下、普遍缺乏医疗保健、缺乏经济发展机遇、政治压迫（尤其针对女性）等。
>
> 暴力冲突：在类似达尔富尔的许多地区，持续不断的冲突每年都会使上千人丧生，数以万计的难民生活在恐惧之中。
>
> 恐怖主义和大规模杀伤性武器：尽管国家间的暴力冲突正在减少，但从历史的角度来看，大规模杀伤性武器可以使较小的国家做出破坏行为，而历史上，只有实力强大的国家才能实施这种破坏。当然，拥有这类武器的小国家也可能会获得惊人的破坏能力。
>
> 全球气候变暖和环境恶化：人类正朝着和平和繁荣的趋势发展，但我们对生存环境的破坏有可能会逆转这一趋势。

这些都是世界范围的问题。和平的国家中也面临很多国内问题，尽管以世界范围和历史的标准来衡量，这些问题微不足道，但这些问题却在一定程度上影响了数以百万计的人，对许多人而言，这些问题甚至关乎生死。正如本书引言所说，在美国存在许多持续不断的争论，涉及税收、医疗保健、移民、平权法案、堕胎、安乐死问题、干细胞研究、死刑、同性恋权利、在公立学校讲授进化论、枪支控制、动物权利、环保规定，以及金融行业管理法规等诸多方面。借助我们对道德心理学的理解，希望我们能够推动这些问题的解决和进步。

本章已经讨论了6种加剧部落冲突的心理学倾向。第一，人们都有部族主义倾向，都会更偏爱“我们”而不是“他们”。第二，关于如何组织社会运行，部落之间存在着真正意义上的分歧，在不同程度上强调了个人权利与更大范围的群体利益。部落价值观在其他层面还有很多区别，比如尊严对于回应威胁的作用等。第三，部落间的道德观各不相同，特别是与宗教有关的道德观。根据宗教道德观，道德权威已经融入了部落的个人、圣书、传统和神灵，即使其他部落不认为这些因素拥有道德权威也没有关系。第四，部族与存在于其中的个人一样，倾向于有偏向性的公平，允许部族层面的利己主义歪曲自己的公平意识。第五，部族思想观念很容易带有偏向性。有偏向性的思想观念源自简单的利己主义，也源自更加复杂的社会动力。一旦思想观念变成了某种文化身份象征，这种观念就能长期存在，即使这种思想观念对部落利益有损也不会改变。最后，就社会问题而言，我们对信息的处理方式会让我们低估自己给他人造成的伤害，最终导致冲突升级。

我们最严重的某些道德问题清晰地诠释了常识道德悲剧——以不同方式展现道德的部落之间的冲突。全球气候变暖或许是最好的例子，哲学家斯蒂芬·加德内尔（stephen Gardiner）将这个场景称为“完美的道德风暴”。第一，全球变暖问题受到有偏见性公平的影响。例如，以限制碳排放的“总量管制和排放交易”制度为例，这一计划限制全球的碳排放，每个国家得到的排放配额有限，可以选择使用，也可以将其转卖他国。关键在于，总量管制和排放交易计划需要对碳排放配额进行初次分配，但在如何公平分配排放配额这一问题上，各国存在明显的分歧。（听上去耳熟吗？）一种方法是基于历史排放水平；另一种方法是根据各国人口，按比例分配排放配额，这样一来，每个人都能得到标准化的碳信用额。如你所想，发达国家倾向于认为，基于历史排放水平的配额分配方案是公平的；发展中国家则更希望按人口比例分配配额。涉及碳排放税收的提案是总量管制和排放交易计划中最受人们青睐的替代物，它要求就碳排放税的征收对象和税率达成一致。目前，全球范围内尚未对碳排放量达成共识，最主要的原因是许多美国人认为1997年减少全球碳排放的《京都议定书》不公平。（美国是全球第二大碳排放国家，绝对排放量和人均排放量都是如此。）乔治·W·

布什在2000年为总统大选拉票时，曾经针对《京都议定书》说过一段代表美国人情绪的话："我将告诉你们，有一件事我绝不会做，我不会让美国像《京都议定书》所规定的那样，背负为全世界净化空气的重担。"不管是美国还是其他国家，至少有一方的公平意识是有偏向性的。

在几乎所有的国际冲突中，定义公平都是一个问题。鉴于巴勒斯坦人的所作所为，以色列占领约旦河西岸公平吗？考虑到以色列人的所作所为，杀害以色列平民对巴勒斯坦人公平吗？使用多少暴力是适当的，多少是过分的？只有某些国家能够拥有核武器，这公平吗？为了抑制独裁统治者，向该国的无辜民众施加严酷的经济制裁是公平的吗？为了使独裁转化为民主，使数以万计的人失去生命是公平的吗？谈判专家和国际关系专家都对冲突解决过程中存在的有偏向的公平长吁短叹。罗杰·费舍尔（roger Fisher）在《基本谈判策略》（*Basic Negotiating Strategy*）一书中对此做出了论述：

> 在谈判中向对手指明，根据我们对公平、历史、道义和道德的理解，他必须做出决定，这种行为最好不过是对眼前任务的偏离，但也可能对我想要的结果造成毁灭性打击。……
>
> 官员们认为自己的行为方式在道德上合情合理。为了让他们转变想法，我们必须唤起他们的是非观念。但这与大多数政府的做法相反。他们首先要唤起民众的是非观念，试图将反对派妖魔化，激起民众支持，这种做法可能会起作用。但随后反对派会变得越来越难对付，……不情愿听我们说的话。

历史学家阿瑟·施莱辛格（arthur schlesinger）在20世纪70年代初期撰写的文章预测出了哈林克谈判研究的结果：对有偏向性公平的追求使每个人的状况变得更加糟糕。

我们放下道德约束，从评判席原谅犯过错误的旧友，这无疑能使我们品行端正的意识得到满足，但却会助长我们对外交政策本质理解的困难。……因为把利益和环境的冲突转化为善与恶的冲突的人必然具有道德优越感。有些人认为外交事务由是非问题组成，他们假定自己比其他人更了解对其他人恰当的事物。他们越是坚信自己是正确的，就越有可能拒绝权宜之计与和解之策，继续寻求道德标准的最终胜利。在国际政治事务中，过度正直几乎是最为有害的行为。

如前所述，从经济不平等到对待堕胎的问题，国内的政治事件也会涉及定义公平的问题。

怎样才能解决新草地上的问题？目前为止，我们已经探讨了道德问题的结构及其背后的心理学特征。与所有动物一样，我们拥有自私的冲动；但与动物不同，我们还拥有社会冲动，自动道德机制迫使我们进入神奇角落，解决了“我”与“我们”之间的对立问题。不幸的是，我们在本章了解到，这种道德机制（伴随着好的、老式的自私与偏见）在更高层面创造出了根本的道德新问题，这些问题出自部落层面，即“我们”与“他们”的矛盾。基于目前的情况，新草地上的问题可能是毫无希望的：社会冲动把我们从个人冲突的泥沼中带出，又将我们投进了水深火热的部落冲突。但值得庆幸的是，人类的大脑可以战胜自私冲动和社会冲动。我们能够思考。为了审视工作中的道德思维，将道德思维和道德情感进行对比，让“心灵”与“大脑”展开竞争的哲学困境便是最好的起点。

# 第二部分 道德反应的快与慢

# 第4章 小火车的学问

本章将会开始介绍我所做研究的核心内容。但首先，我将对自己选择这项研究的缘由以及本研究的重要性进行简单介绍。

我上八年级时，加入了学校的辩论队。我所参加的是林肯–道格拉斯辩论，要求两名辩手就同一“决议”以不同立场进行辩论。决议由国家委员会制定，每过几个月便会更改一次。最近，我在网上找到了自己年少无知时曾讨论过的题目，下面列出的便是我读高二时一些可供辩论的题目：

·决议如下：与国内问题相比，美国应当更加重视全球问题。

·决议如下：美国全体国民都应服兵役。

·决议如下：美国的社区团体应当获得打击黄色书刊的权利。

·决议如下：发展自然资源比环境保护更加重要。

·决议如下：确保道德的公共服务，法律对个人的约束比良知的约束更加有效。

当时的我尚未意识到，这些问题都是我们身边关于合作的问题，有些是关于社会中个人的合作，有些则是关于国家之间的合作。

夜晚和周末的高中教室空旷无人，我和队友穿着不合身的里根–撒切尔风格的套装鱼贯而入，再现了新草地上的哲学纷争。尽管我们的立场是随机分配而来的，我们依然充满激情地为己方立场辩护。很快，我建立了一种标准辩论策略。开始阶段，每位辩手都会提出一个“价值前提”，即己方最重要的价值观。例如，假如你要反驳对黄色书刊的打击，那么“自由”可能就是你的价值前提。如果你认为法律约束比良知约束更加有效，那么“安全”便可能成为己方的价值前提。随后，辩手会论证己方价值前提的重要性。例如，如果“安全”是你方的价值前提，你可能会引用托马斯·霍布斯的一些名言，证明安全是最重要的考虑，因为安全是其他价值观得以落实的前提条件。最后，你便可以基于己方的价值前提，证明己方观点更加符合该价值前提。

我并不喜欢这些标准的价值前提（“自由”、“安全”等），因为在我看来，不论你最看重何种价值，总有一些更加重要的因素需要考虑。的确，自由

非常重要，但自由可以囊括一切吗？安全也确实十分重要，但安全就可以囊括一切吗？如何才能找到最为重要的价值观呢？这时，我发现了功利主义，这是18世纪和19世纪的英国哲学家杰里米·边沁（Jeremy bentham）和约翰·斯图亚特·密尔（John stuart Mill）探索出的哲学观点。*

功利主义虽然名字不好，却是一个伟大的思想。在我看来，功利主义是一切道德哲学和政治哲学中被低估最多，也承受最多误解的学说。从第三部分到第五部分，我将解释功利主义缘何如此智慧，它又为何饱受误解，不被看好。但为了使我们更加易于接受功利主义的心理机制，我们首先从一个简单积极的例子入手。读完本章后，你对于功利主义的感觉可能会五味杂陈，也可能会更加糟糕，但这都没关系。在本书的第三、四、五部分，我会努力将你争取过来。

功利主义的整体思路是什么呢？功利主义认为，我们应该做的是能够将所有人的共同利益最大化的行为。（严格说来，这里描述的其实是结果主义，这是一个更加广泛的哲学范畴，包含功利主义的概念。第六章会对此进行更多介绍。）换句话说，我们应当尽可能地为多数人谋取福利。比如，若选择a会杀死6个人，拯救4个人；选择b会杀死4个人，拯救6个人，假设所有其他的后果都完全相同，那么我们就应选择b。这个想法可能会使你感到震撼，很明显，我们甚至无法将其称为一种“思想”，更不用提什么“伟大”了。但我们将很快意识到，宏观考虑道德问题时，这种思考方式是很好的切入点，这一点并非显而易见。同时，我们在第三、四、五章将会提到，在混乱的现实世界应用这项原则并非易事，其实际应用与人们想象中的功利主义应用也是大相径庭。

作为一名辩手，我十分偏爱功利主义，因为功利主义作为一种价值前提，其本身就能够平衡不同的价值观：究竟是自由更重要，还是安全更重要？功利主义给出了理智的回答：没有绝对意义上的更重要，我们要在自由和安全之间找到平衡，最好的平衡点就是能创造出最大公共福利的选择。

我对这个策略十分满意，不论立场如何，我在每场辩论中都将功利主义作为价值前提。每场比赛，我以对功利主义的高谈阔论作为开场，中间夹杂着引用一些密尔及其朋友的权威论断。从那以后，我所要做的只是选出最好的证据，证明随机分配给我的观点能够为多数人的利益服务。

这项策略非常好用。我不仅将功利主义作为己方价值前提，还将其作为武器，攻击对方的价值前提。不论对方提出何种价值前提，我都会采用最夸张的方式，将他们的价值观与多数人的利益进行比较。我通常会以盘问（盘问环节中，双方辩手相互直接发问）而不是陈述的方式开局。例如，如果我的对手维护言论自由，我就会抛出早已准备好的反面例证：

我：你认为本场比赛中，言论自由是最重要的价值观，是这样吗？

不明就里的对手：是的。

我：因此，任何价值观的重要性都不会超过言论自由，是这样吗？

不明就里的对手：是的。

我：那么……假设有一个人，为了好玩，在拥挤的剧场里大喊“着火了！”，人们慌忙中冲向出口，有人因被踩踏而丧生。那么随意喊“着火了！”的权利是否比不被踩踏致死的权利更加重要呢？

一针见血！反驳“言论自由”并不困难，“拥挤的剧场”是一个古老且好用的例子。但多数情况下，我都能够想出或编出一个不同的反面例子。

18世纪的德国哲学家伊曼努尔·康德（immanuel Kant）被很多人（特别是功利主义的批判者）誉为有史以来最伟大的道德哲学家。在辩论过程中，辩手也对康德喜爱有加，我的对手有时会援引康德的“绝对命令”作为己方价值前提。他们会说，“不能通过结果来判断手段的好坏”。这时我便会提出质询：“假设一台坏了的电梯就要砸到人了，能够阻止惨剧的按钮是你所够不到的，但你可以将一个人推向按钮。为了拯救一个人的生命，将另一个人作为按按钮的工具，你觉得这样做合理吗？”

我喜欢将功利主义作为辩论策略，并非因为它好用，而是因为我信任它。当然，就像刚才说的那样，不论我的观点是正是反，我都需要论证己方立场符合多数人的利益。但这种论证不过是游戏的一部分而已，对我来说，与其每场辩论都为己方立场提出并不完善的理论，这种论证过程似乎更好一些。

一次，我参加了在佛罗里达州杰克逊维尔市举办的比赛。我的对手来自迈阿密，是一位非常尖锐的辩手。按照惯例，我流利地陈述了功利主义的原理，然后接受她的质询。

我的对手：你认为我们应当尽己所能照顾大多数人的利益，是这样吗？

不明就里的我：是的。

我的对手：那么……假设现在有5位病人因器官衰竭，濒临死亡。这5个人衰竭的器官各不相同，一人病在肝脏，一人病在肾脏……

不明就里的我：嗯。

我的对手：假设有一位拥护功利主义的医生，他可以绑架一个人，将其麻醉，把他的器官取出，分配给另外5位病人。这样似乎能够满足多数人的利益。你认为这样做对吗？

我震惊了。我已经不记得自己当时是如何回答的了。也许我将功利主义引到了现实中，声称不道德的器官移植并不会满足多数人的利益，因为人们会因此担心器官移植被滥用，因而生活在恐惧之中，等等。但不论我说了什么，都不足以挽回局面，我输掉了那场比赛。更糟糕的是，我失去了自己的制胜法宝。

最糟糕的是，我那正在茁壮发展的道德世界观遭遇了危机。（对于没有女朋友的十几岁男孩子来说，失去茁壮发展的道德世界观可以说是损失巨大！）

在高中三年级读到一半时，我接到了大学的录取通知，随后便退出了辩论队。我的父母为此十分难过，辩论队教练将我叫作叛徒，但那时的我已经厌倦了比赛形式的辩论。如果一定要提出哲学论点，我希望自己真心信服这个观点。但那时候，我并不知道自己信服什么。

1992年秋，我以本科大一新生的身份进入了宾夕法尼亚大学沃顿商学院。我花了一个月的时间才意识到商学并不适合自己，但在第一年的学习中，我遇到了几位教授，接触到了一些观点，他们对我产生的影响一直持续至今。第一学期，我在学习微观经济学的过程中接触到了博弈论。博弈论研究的是囚徒困境和公地悲剧等状态下的战略决策问题，这项理论的抽象和优雅令我深深着迷。全球变暖、核武器扩散、宿舍内公共厨房永远的脏乱，许多看似不相关的社会问题背后，都有着相同的数学结构支撑。只要我们能理解这些问题的数学本质，就能找到解决方案，这个想法也令我十分着迷。

那一年，我去听了人生中的第一堂心理学课，遇到了保罗·罗津，一位杰出的教师兼科学家。与通常情况下在大礼堂中开设的入门课程不同，这门课以小型讨论会的形式进行。罗津会向我们提问，与我们争辩，引导我们进行数学证明。有一次，我们一同计算了人类神经系统传送电信号的速度。我们使用的是德国物理学家、外科医生赫尔曼·冯·亥姆霍兹（Hermann von Helmholtz）发明的算法。他是19世纪实验心理学的奠基人之一。

首先，罗津让我们手牵手连成一队，让第一个人捏一下第二个人的手，第二个人再捏第三个人，依此类推。“捏”这个动作需要多长时间才能传到队

尾呢？我们反复进行了多次试验，罗津握着秒表，记录每次试验所需的时间，并算出平均值。接下来，我们再次重复试验，但这次试验中每人都握住下一个人的脚踝。当我们感到左脚脚踝被捏了一下时，就立即用右手捏下一个人的左脚，直到队尾。同样，罗津要求我们反复试验多次，算出动作传到队尾的平均用时。这次试验的平均用时比上次要长一些。我们还测量了手和脚踝分别到大脑之间的距离。两者的差值就是“捏”这个动作在第二次试验中需要额外传输的距离。通过计算多次试验的平均值，我们估算出了神经信号在这段额外的距离上传输所花费的时间，从而估算出它的传输速度。罗津事先背着我们写下了教科书中给出的答案，而我们的计算几乎与书上的答案完全一致。

类似的数学证明让我对科学思维的力量刮目相看。具体说来，这些试验向我展示，人类思维的奥秘能够以巧妙的方式揭开，转化为有着确切答案的问题。令我着迷的另一个原因是，早在几千年前，人们便有条件进行这样的试验，但直到19世纪，才有人想到这样的方法。（如果参与者足够多，试验次数也足够多，甚至连秒表都不需要。）在这里，先进的技术并不重要，将强大的推理能力和创新能力相结合的思维方式才是力量之源。

罗津还向我打开了生物社会学的大门。生物社会学从进化的角度研究社会行为，特别是人类的社会行为。（该学科现在包括几个分支，进化心理学便是其中之一，主要研究进化过程对人类思维的影响。）生物社会学家能够解释为何女生（请原谅我使用大学中如此原始的称呼）只要想，就能随时与人上床，但男生却不行。罗伯特·泰弗士（robert Trivers）的亲代投资理论可以解释这一现象：女性为了繁育出成熟的下一代，需要投入大量精力——九月怀胎，数年哺乳。而男性所付出的不过是一团精子而已。（当然，在后代身上投入更多精力的男性也更有可能养育出成功的后代，但男性的最低投入相对较少。）因此，泰弗士认为，女性在选择另一半时会更加慎重。如果某位女性不加选择地接受第一位男性的精子，她的后代可能不如慎重选择女性的后代身体健康。但对于男性来说，情况则正相反，他们可以更加自由地献出自己的遗传物质，而不必付出任何代价。［这使我想起一首当时非常流行的歌曲：红辣椒乐队的Give it away（送出去）。］很多人对这种观点不屑一顾，认为这是用伪科学的手段论证传统性别角色，但我却为此深感震撼。我并不是传统性别角色的拥护者，恰恰相反，如果我周围的女性没有那么挑剔，我会十分开心的。事实上，泰弗士的理论不仅对显性的社会现实做出了解释，而且还正确预测出了一些不甚明显的现象。泰弗士认为，男女性别本身的对立并不重要，重要的是双方的亲代投资差距很大。如果某个物种中的雄性是生育过程中投入更多的一方，那么慎重选择另一半的便会是雄性而不是雌性。对某些鸟类和鱼类来说，保卫和养育幼体的责任由雄性承担，这些动物中择偶更加慎重的一方果然是雄性。

大学中，我第一次自己掌握财政大权，与自由相伴而生的是新的责任感。每月的零花钱大多花在了小件奢侈品上：音像店里买来的CD（激光唱片），费城中心城新鲜的小吃等。我是如何说服自己的呢？我幻想有一位绝望而穷苦的女人向我哀求："请给我10美元好吗？没有钱就买不起吃的，我的孩子会饿死的。"我忍心看着她的眼睛拒绝她吗？难道我忍心说："很抱歉，我需要一张约翰·柯川的CD。你的孩子大概只能饿死了。"尽管我从未在流浪汉身边停下脚步，但我知道我永远做不出这样的事。然而与此同时，我看到了这种争论的结果：世上有太多人深陷绝望，他们对钱的需求远大于我对于更多音乐CD 的需求，那么我的责任何时才能终止？（第三部分中会详细介绍这个问题。）

我找到了一位心理学教授，乔纳森·伯龙（Jonathan baron）。他的简历显示，他对心理学、经济学和伦理学都很感兴趣，看上去像是可以与之对话的类型。我通过预约与他见面，他说，我提出的问题也是哲学家们一直以来争论的话题。这个问题由彼得·辛格（Peter singer）首次提出，那时我还没有出生，从那以后，争论从未停止。他自己也被这个问题困扰许久。

伯龙和我是一对极好的搭档，我们开始共同进行研究。那时候，我认为每所大学都拥有这样的教授，但直到后来，我才意识到，那时的我虽然充满激情，但能够遇到世界上唯一的一位正牌功利主义道德心理学家，我是多么幸运。伯龙和我研究了环境决策中的"忽视数量"问题。如果你向别人询问，"为了清洁两条受到污染的小河，你愿意支付多少钱？"你会得到一个答案。如果你向另外一些人询问，"为了清洁20条受到污染的小河，你愿意支付多少钱？"你所获得的答案与第一次几乎相同。你可能会认为，或者说你希望人们心中认为，清洁20条小河会比清洁两条小河好10倍。但通常情况下，数字并不重要。不论是两条小河，还是20条小河，听上去都是一样的。*伯龙和我试着弄清人们为什么会"忽视数量"。我们的研究并没有解决这个问题，但我们确实排除了一些可能的理论，这就已经是一点进步了。这个研究项目帮助我发表了第一篇学术论文。但更重要的是，与伯龙一起的研究经历让我接触到了对"探索与偏见"的研究。探索是人们做出决定的思维捷径，而偏见则是探索性思维所导致的非理性过失。

我对商学失去兴趣后，便转到了哈佛大学的哲学专业。在哈佛大学的第一个学期，我去上了一门叫作"思考思维"的课程。这门课程由哲学家罗伯特·诺齐克（robert Nozick）、进化生物学家斯蒂芬·杰·古尔德（stephen Jay Gould）以及法学教授艾伦·德肖维茨（alan Dershowitz）共同执教。三位都是传奇式的教授，这门课也因此被戏称为"名人自我"。课程纲要上印有哲学家朱迪斯·贾维斯·汤姆逊（Judith Jarvis Thomson）的一篇论文，题目为《小火车问题》。

## 小火车问题

我发现，高中时横空出世将我击败的器官移植困境便源自这篇论文。这篇论文观点新颖，讨论了一系列的道德困境，每种困境都源于一命换五命的话题。与器官移植困境相似的一些例子是明显错误的。但论文中还有另外一类例子，如人行天桥困境，我稍作修改，摘录在下面：

一列失控的小火车冲向了五位铁路工人，照这种情形发展下去，五人都会丧命。你站在一架横跨道路的人行天桥上，刚好位于小火车和五位铁路工人中间。你身边有一位铁路工人，背着一个双肩大包。挽救五人生命的唯一方法就是将身边的工人推下天桥，落到铁轨上。这个人当然会死去，但他的身体和双肩大包会挡住小火车，保护其他人。（你不能选择自己跳下天桥，因为你没有双肩大包，无法挡住小火车，你也没有再背一个包的时间。）为了拯救五个人的生命，将身边的陌生人推向死亡，在道德层面是可以的吗？（见图4.1）

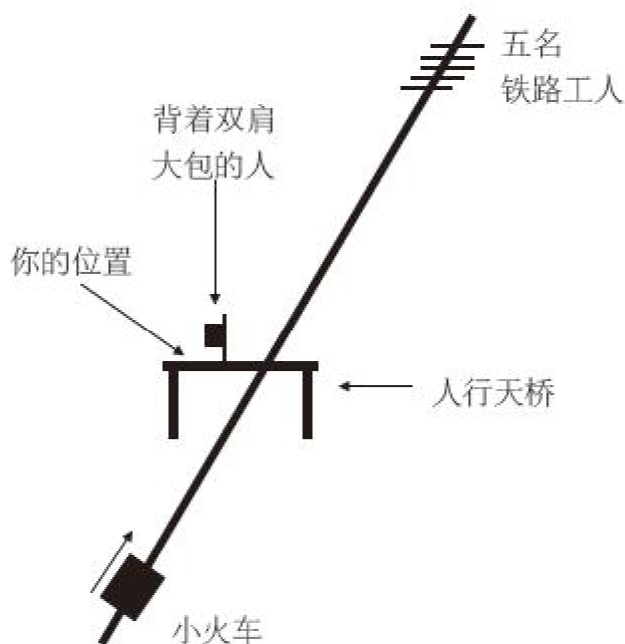



图4.1 人行天桥困境

大部分人认为不能为了拯救另外五人而将一个人推下天桥。然而，如果我们接受这个困境的所有前提，这个答案并不符合功利主义的思路。尽管把这个人推下天桥能够使多数人获利，但这种做法似乎并不正确。

要想摆脱这种困境，我们可以做出很多尝试。其中最诱人的方法就是质疑人行天桥困境的前提：将身边的工人推下天桥真的能够挽救五人的生命吗？是否存在其他方式可以挽救五人的生命？如果推人下桥这件事刚好被他人看到，那个人由此不再珍视人类生命，最终沦为杀人犯怎么办？如果这种杀人行为被容许，成千上万人的生活都会因这种功利主义行为而蒙上阴影，这该怎么办？这些都是非常合理的问题，但它们并无益于解决这个问题。在更加现实的假设下，功利主义思维也可能找到合理的理由，反对将人推下天桥。这是非常重要的一点，我稍后会加以强调。但眼下我们需要暂时搁置怀疑，认真判断，即使将这个人推下天桥能够使多数人获利，这种做法也是不正确的。

为什么呢？对功利主义最常见的诟病就是对人类权利的低估，因为功利主义允许我们不考虑后果，对他人做出完全错误的行为。我在前文中提到了康德的绝对命令，他自己将这个概念总结如下：

在行动时，要把你人格中的人性和其他人人格中的人性，在任何时候都看作目的，永远不能将其只看作手段。

Act so that you treat humanity， whether in your own person or that of another， always as an end and never as a means only.

粗略翻译一下，意思就是：不要利用他人。利用他人阻挡火车，已经是最为极端的情况了。

对人行天桥困境来说，好的一方面是很多其他命题由此衍生出来。其中一个变种被称为开关困境：一列失控的小火车冲向了五位铁路工人，照这种情形发展下去，五人都会丧命。你可以扳动一个开关，将小火车引上另一条支线，从而挽救5人的生命。但不幸的是，支线铁轨上也有一位铁路工人，如果你扳动开关，这个人便会因此丧生。（见图4.2）

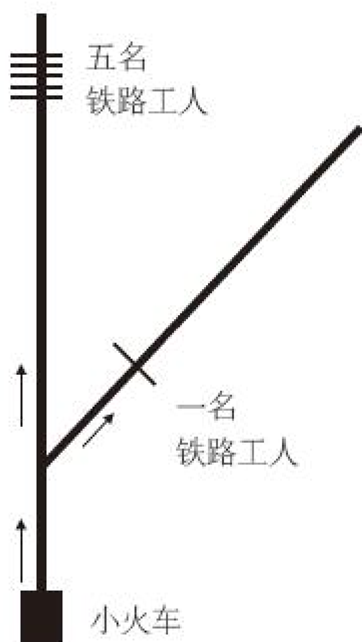



图4.2 开关困境

让小火车避开五位工人，撞上一位工人，在道德层面是可以的吗？对于汤姆逊来说，这种做法似乎更易接受，我也同意他的观点。我们在后面将会发现，全球各地的人们都同意这个观点。那么为什么我们在开关困境中选择同意，而在人行天桥困境中选择反对呢？

在我看来，这是一个完美的科学问题。从10岁以来一直困扰我的所有问题都包含在了小火车问题这个美丽如果蝇的模型当中。首先，小火车问题讨论的是隐藏在高中学校所有辩论话题背后的深层哲学问题，并对问题的本质进行了归纳：在何种情况下，个人权利高于多数人的利益？为什么？很多主要道德问题，包括堕胎、平权措施、税收的高低、战争中平民的死亡、将人们送上战场、医疗服务中资源的分配、枪支管理、死刑等，或多或少都是个人权利（不论是真实的还是名义上的）与多数人利益（不论是真实的还是名义上的）之间的对抗。小火车问题则正中靶心。在人行天桥困境中，为了多数人的利益而牺牲一个人的生命似乎是不对的，是对人权的公然践踏。但在开关困境中，用一人的牺牲来换取另外五人的得救虽然远非理想方案，但这种做法却看似合理。在同一个小而精的谜题中，出现了康德与密尔之间的对抗。如果我能想通这两个简单的困境，我就能明白更多的道理。

小火车问题中存在一种美感，一种亥姆霍兹式的简洁。如何计算神经系统中信号的传输速度？你不必跟随信号向上传入胳膊，穿过大脑复杂的迷宫，再向下传入另一条胳膊。只需将传输媒介由胳膊换为腿，然后做减法计算就可以了。小火车问题也把我们引向了可爱的减法计算，不同困境的变种也许拥有很多潜在的重要特点，但开关困境和人行天桥困境之间的差别却并不很大。在仅有的几个差别中，一定有一个与道德选择相关，至少事实看似如此。

小火车问题也是一个决策问题，从“探索与偏见”的角度来看就可能会有所收获。直觉告诉我们，人行天桥困境中的行为是错误的。那么，大脑的何种机制使我们得出了这样的结论呢？有时候人类的大脑认为清洁20条被污染的小河与清洁两条被污染的小河并没有太大差别。在我们本应做出不同反应时，我们的反应却并无不同。小火车问题体现的则是恰好相反的偏见：用截然不同的态度对待相似的事情。一命换五命的做法在一个例子里是正确的，在另一个例子里就变成了错误的，这也许只是人类古怪的心理所导致的。这是一个诱人的想法，它似乎能够维护功利主义观点。因为若非如此，功利主义观点似乎过于不切实际。

“思考思维”课程结束后的那个夏天，我获得了一笔并不丰厚的经费，对小火车问题进行独立研究。我阅读了大量的哲学书籍和心理学论著，为了履行获得经费的义务，我写了一篇题为《两种道德》的论文。文中指出了两种不同的道德思维方式，分别被称为“抽象的”和“共鸣的”。这就是道德判断中“双加工”理论的开始，我稍后会对此详加阐释。

第二年春天，我选择了行为神经学的课程，希望能够从研究大脑的学者那里找到答案。课堂之上，我并没有找到答案，但却偶然发现了一本新近出版的著作：《笛卡儿的错误》（*Descartes’ Error*）一书由神经学家安东尼奥·达马西奥（antonio Damasio）所著，讨论了情感对决策的影响。达马西奥介绍了菲尼斯·盖奇的著名案例：19世纪时，盖奇在佛蒙特州生活并工作，他是一位受人尊敬的铁路领班。一次意外爆炸事故中，一根三英尺长的铁条插入了他的眼眶，从头骨顶部穿出，从那以后，盖奇的性情大变。这次意外使盖奇的内侧前额叶皮质大面积受损，受伤部位为眼睛和前额后部的大脑，就在鼻子上面。不可思议的是，几周之后，盖奇似乎恢复了认知能力，他能够说话、做数学题，还能记住人名和地名。但盖奇再也不是从前的盖奇了，从勤奋努力的铁路领班变成了不负责任、游手好闲的人。

达马西奥对内侧前额叶皮质腹侧（腹内侧前额叶皮层，VMPFC）受损的病人进行研究时，发现了一个规律。这些与盖奇情况类似的病人在智商测试等标准认知测试中表现良好，但在现实生活中，他们做出的决定却很糟糕。在一系列的研究中，达马西奥和同事发现，病人的问题来源于情感的缺失。在血淋淋的车祸照片和溺水而亡的洪灾难民照片前，一位病人反馈

说，自己对这些悲惨的场景感到无动于衷，但他知道，脑部受损之前，自己对这类情景会做出非常情绪化的反应。达马西奥将病人的窘境描述为“能够知道，却不能感觉”。

读这本书时，我正独自坐在宾馆的房间里，看到这段话，我一下子兴奋起来，站到了宾馆的床上，起劲儿地跳了起来。这段话与小火车问题之间的联系让我灵光顿现：这些病人所缺失的能力恰好就是决定普通人对人行天桥困境态度的机制。当然，我们完全可以对这个想法进行检验。我们可以让腹内侧前额叶皮层受损的病人在开关困境和人行天桥困境中做出选择。如果我的想法是正确的，那么在两种困境中，这些与菲尼斯·盖奇情况相似的病人均会做出符合功利主义思维的选择。但不幸的是，我不认识这样的病人。

接下来的一年中，我将这些想法写进了自己的毕业论文，题为《道德心理学和道德进步》，这篇论文也是这本书早期的一个雏形。1997年秋天，我被普林斯顿大学录取，攻读哲学博士。读博士的前两年，我忙于参加讨论课，为了让自己达到各项要求，我涉猎了不同课题，从柏拉图的《理想国》到量子力学的原理都有所了解。总体来说，我十分享受作为哲学家的生活。1999年夏天，我得知一位心理学系的神经学家想要找一位哲学家谈话。普林斯顿大学邀请乔纳森·科恩（Jonathan Cohen）出任新成立的“大脑、心理、行为研究中心”的负责人。我查阅了他的网站，发现他在研究过程中使用了脑成像技术。我想，也许我不再需要神经受损的病人了，也许我们可以在健康人的大脑中直观地看到人行天桥困境对人产生的影响。于是，我与他预约见面。

科恩的实验室里堆满了没来得及打包的箱子，办公室里的书和论文堆得摇摇欲坠，像是“学术钟乳石”。他靠在办公室的椅背上，说：“那么，你现在有什么想法？”我开始向他解释小火车问题，从开关困境说到人行天桥困境。他将我打断，列出了两种困境的10处不同。“等一下。”我说道，继续描述达马西奥的著作和菲尼斯·盖奇的例子。还没等我说完，他便冲口嚷道，“我知道！我知道！我知道了！腹侧背侧！腹侧背侧！”我知道“腹侧”这个词，却不明白“背侧”指的是什么。是像鲨鱼鳍那样的吗？（背侧指的是大脑的上半部分，与四足动物的背部平齐的部分。）但不管怎样，他对此感到兴奋，我已经很满足了。“我们可以一起做，”他说，“但你必须学习脑成像技术。”这对我来说似乎不错。

这次碰面中，科恩想到的也是我那时没能理解的，是小火车问题作为神经学案例的另外一半故事，而这一半故事也是与科恩的工作联系最为密切的部分。之前，我一直思考的是，情感对人们在人行天桥困境中否定的选择

起到了怎样的作用。但我们为何又会在开关困境中做出肯定的选择？对菲尼斯·盖奇和其他情况类似的人来说，他们大脑中未受损的部位又是哪些？如果这些病人能够“知道”，却不能“感觉”，那么使他们“知道”的，又是什么机制？对我来说，答案十分简单：这一切都是由功利主义的成本效益分析思维导致的，也就是说，挽救五条生命优于挽救一条生命。但是对于认知神经学家来说，与思维机制有关的一切问题都绝非简单。

科恩是认知控制神经学实验室的负责人，根据他的定义，认知控制是“将思维与行动按照内心中的目标协调一致的能力”。认知控制能力的一个经典测试就是斯特鲁普辨色任务，要求参试者辨认屏幕上出现的字体颜色。比如，你可能会看到用蓝色字体显示的“小鸟”一词，那么你的任务就是说出“蓝色”。但如果这个词本身就是描述颜色的词，而这种颜色与其显示字体的颜色又不相符，这时的任务就不那么简单了。比如，看到用绿色字体显示的“红色”一词，你的任务是说出“绿色”，但你的第一反应也许是“红色”。因为与辨认颜色相比，阅读是更加无意识的反应。这些任务反映了大脑内部的矛盾，一部分神经元在说，“读出词语！”另一部分神经元则说，“说出颜色！”（当然，神经元是不会说话的，为了便于解释，我采用了拟人的修辞。后文也会有类似的情况。）是什么将这些相互矛盾的命令协调一致？又是什么能够保证最终协调结果正确（“说出颜色”），而不是出现错误（“读出词语”）？

这就是认知控制，它能够反映人类的认知能力，由背外侧前额叶皮层（DlPFC）的神经回路控制。在斯特鲁普辨色任务中，背外侧前额叶皮层会说：“嗨，各位，我们现在的任务是辨认颜色。所以，辨认颜色的神经元们，请活跃起来；阅读词语的神经元们，请安静下来。”背外侧前额叶皮层能够用清晰的决策标准（“说出颜色”）引导人类行为，并且能够压制与其相反的冲动（“读出词语”）。科恩喊出的“腹侧背侧”，便是这个意思。作为一名认知控制神经学专家，科恩立即意识到，“拯救更多人”与“说出颜色”一样，都是一项清晰的决策标准，能够决定我们对问题的反应态度。此外，科恩还意识到，在人行天桥困境中，将人推下天桥以拯救更多性命的功利主义思维就好比是在斯特鲁普辨色任务中，辨认用绿色显示的词语“红色”的颜色。要想给出符合功利主义思维的答案，就必须压制相反的冲动。

将所有概念整合起来，就得到了道德判断的“双加工”理论。我们可以通过人行天桥困境和开关困境加以解释：之所以叫作双加工理论，是因为这种机制预设了两种不同甚至相互矛盾的反应，两种反应都是自发的、受到控制的。（下一章会对双加工理论做更多介绍。）在开关困境中，我们有意识地利用背外侧前额叶皮层，将功利主义作为决策标准。这个场景中，会造成伤害的行为并没有引发强烈的情绪反应（具体的原因我们稍后再做讨

论），因此我们倾向于做出符合功利主义思维的选择，即扳动开关，使更多的人获救（参见图4.3上部）。在人行天桥困境中，我们也利用背外侧前额叶皮层，将功利主义作为决策标准。但在这个场景中，出于某些原因，会造成伤害的行为触发了相对强烈的情绪反应。这种反应来自腹内侧前额皮质层（VMPFC）。因此，尽管与功利主义的成本效益分析思维背道而驰，多数人依然认为一命换五命的行为是不正确的。（参见图4.3下部）这就是我们想要证明的理论。

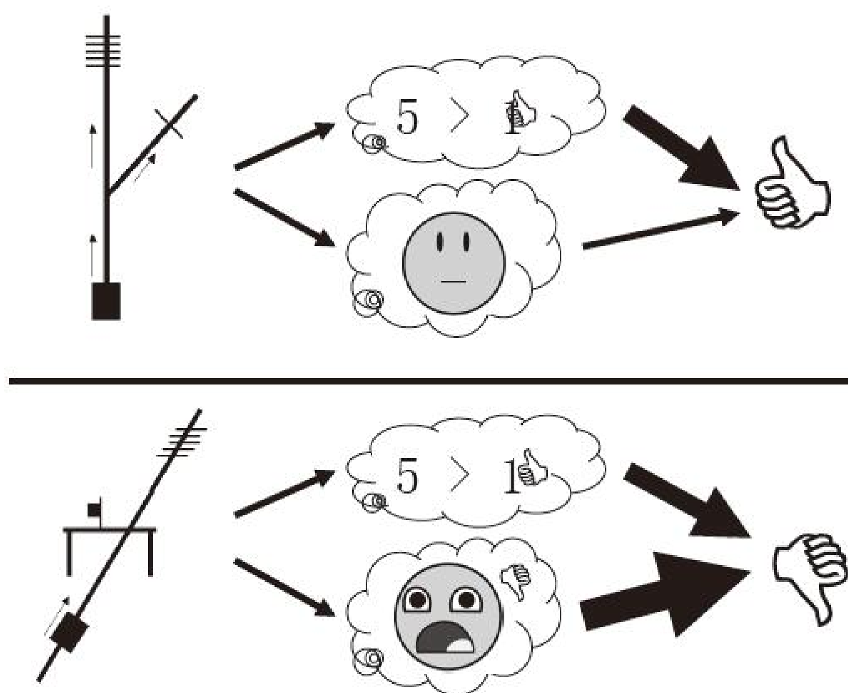



图4.3 道德的双加工理论

让小火车避开5 个人而撞向一个人（上图）符合功利主义思维，而且没有触发强烈的抵触情绪，因此多数人持赞同意见。将一个人从人行天桥推下去（下图）也符合功利主义思维，但与此同时，强烈的负面情绪被触发，因此多数人持反对意见。

## 扫描大脑

第一次试验中，我们设计了一系列道德困境，一些与开关困境类似，另一些则与人行天桥困境类似。我们将两组情景分别称为“客观的”和“主观

的”。参试者需要阅读两组情景并做出选择，与此同时，我们用功能性磁共振成像技术对参试者的大脑进行扫描。**随后，我们对人们面对两类不同道德判断时的脑成像数据进行分析，按照亥姆霍兹的试验思路，将两组数据进行对比，用减法进行分析。正如我们所料，在类似人行天桥困境的“主观的”情景中，内侧前额叶皮质，包括腹内侧前额叶皮质的活动都更加活跃。*也就是说，与人行天桥困境相类似的情景所触发的大脑活动区域恰好就是菲尼斯·盖奇的脑部受损区域，也是达马西奥的案例中，那些能够“知道”，却不能“感觉”的病人的脑部受损区域。与此相反，与开关困境类似的“客观的”情景中，背外侧前额叶皮层的活动明显增强。之前科恩使用斯特鲁普辨色任务作为场景进行脑成像试验时，已经多次观察到了同一区域的活动增强。第二次试验结果显示，面对与人行天桥困境相似的场景，如果人们的选择符合功利主义思维（例如，赞成以一命换五命的选择），背外侧前额叶皮层附近某个区域的大脑活动会显著增强。试验结果还显示，与“客观的”、类似开关困境的场景相比，当人们对“主观的”、类似人行天桥的场景进行思考时，大脑的另外一个区域——杏仁核也会变得更加活跃。众所周知，杏仁核对人类的情感变化有着重要的影响。也许你还记得，在第2章中，杏仁核的活动与外族人的辨认和警惕性的提高有关。

这些发现与我们的假设完全吻合，我们感到十分惊喜。但整个试验依然存在值得商榷之处。首先，我们的试验并没能像预想的那样实现严格控制。在理想状况下，试验中只使用两个场景：开关困境和人行天桥困境。由于两个场景十分相似，便能够排除更多的干扰因素。但脑成像数据过于复杂，仅对两个独立事件进行数据比较十分困难。（同样的道理，罗津让我们手牵手反复试验多次，手握脚踝的试验也要反复进行多次，然后求取平均值，以此获得大致的速度差。）这就意味着，我们需要很多与人行天桥困境类似的例子和很多与开关困境类似的例子进行比较。但我们并不能使两类场景过于相似，否则人们可能会放弃思考，直接给出相同的答案。更糟糕的是，我们并不知道人行天桥困境和开关困境的关键区别在哪里。因此，要想设计一系列与人行天桥困境相似的场景和一系列与开关困境相似的场景，是十分困难的。同类别的场景中，如何界定相似？（在后来的研究中，我们试图解决这个问题，第9章对此问题有所介绍。）我们只能根据猜测确定两类场景中的关键区别。我们知道这种猜测未必正确，但依然希望我们的猜测能足够接近真相，使我们能证实关于“腹侧”和“背侧”的假设。结果是，我们成功了。

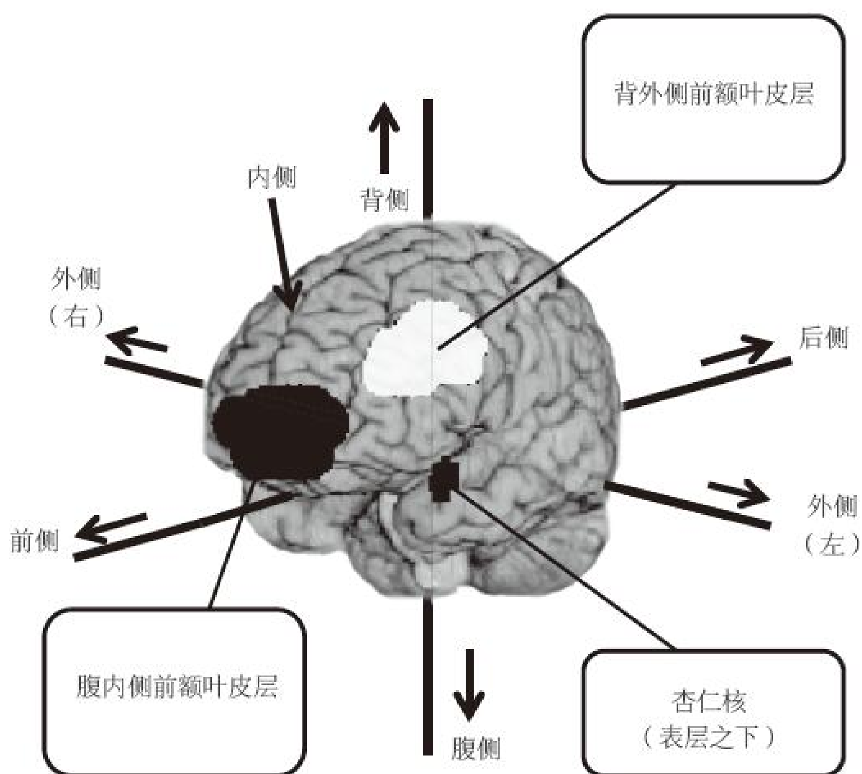



图4.4 三维脑成像图，标出了道德评判中涉及的三个大脑区域

我们的初步试验还有一个重大局限：脑成像数据是“相关性”数据，也就是说，我们无法确定图像中反映的大脑活动是否促成了人们的判断，或是，图像中反映的大脑活动是否仅与人们的判断相关。例如，冰激凌的销售量与溺水事件具有相关性，但冰激凌并不是导致溺水事件的原因。事实是，在炎热的夏天，人们会买更多的冰激凌，也更愿意去游泳，而游泳才是溺水事件的诱因。正如科学家们常说的那样，“相关性不能反映因果关系。”（但很多科学家过分讲求实际，因而常常忽略的一点是，相关性可以作为研究因果关系的切入点。）因此，思考“主观的”困境问题与腹内侧前额叶皮层以及杏仁核的活动具有相关性，但这些区域的大脑活动是否导致人们在与人行天桥困境类似的场景中做出否定选择呢？同样的，思考“客观的”困境问题与背外侧前额叶皮层的大脑活动相关，背外侧前额叶皮层的大脑活动与在“主观的”场景中做出符合功利主义思想的选择也相关，但背外侧前额叶皮层真的是导致人们在开关困境，以及少数的人行天桥困境中给出肯定选择的原因吗？

## 试验型小火车学的开端

科学的乐趣之一在于，你可以向世界抛出一个观点，然后看着其他科学家对其进行研究。我们完成前两个脑成像试验后，很多不同领域的研究人员使用各种不同的方法，在我们的研究基础上继续深入，为我们的理论提供了新的证据，也提出了新的研究方向。当然，我们自己也做了一些后续研究。*

接下来一项重大的小火车研究来自于加州大学洛杉矶分校的马里奥·门德斯（Mario Mendez）教授和他的同事。他们对额颞叶痴呆（FTD）病人的道德判断进行分析。额颞叶痴呆是一种进行性神经疾病，会对腹内侧前额叶皮层造成影响，因此，额颞叶痴呆病人往往会出现行为异常，与菲尼斯·盖奇等腹内侧前额叶皮层受损的病人十分相似。特别值得注意的是，额颞叶痴呆病人的突出特点就是“情感迟钝”和缺乏同理心。门德斯和他的同事将不同版本的开关困境和人行天桥困境分别展示给额颞叶痴呆病人、老年痴呆病人和健康人，试验结果与我们的预测惊人的一致。在开关困境中，三组参试者的选择分布完全相同，每组都有80%以上的人赞同扳动开关，让小火车变轨以挽救五条性命。在人行天桥困境中，大约20%的老年痴呆病人赞同将人推下天桥。以健康人作为对照组，结果也大抵相同，但额颞叶痴呆病人中却有60%都赞同将人推下天桥的行为，这一数字是其他两组的三倍。

这项研究解决了之前提到的两个问题。首先，由于研究人员直接对病人进行测试，而不是使用脑成像技术，因此并不需要设计很多类似的困境来计算平均值。这就避免了对“主观”（与人行天桥困境类似的）和“客观”（与开关困境类似的）进行定义的问题。研究人员只需比较最初匹配的两个困境即可。其次，这项研究解决了相关性和因果关系的问题。也就是说，研究人员更加确定地验证了：情绪反应是导致人们在人行天桥困境中做出否定选择的原因。因为试验结果显示，有情感缺陷的人做出肯定选择的概率是其他人的三倍。

几年之后，达马西奥亲自将我在宾馆的床上蹦跳时所想象的试验变成了现实。在迈克尔·柯尼希斯（Michael Koenigs）和利亚纳·杨（liane Young）的带领下，达马西奥与自己的搭档将我们设计的全套困境场景展示给像菲尼斯·盖奇一样腹内侧前额叶皮层受损的患者。毫无悬念，这些患者在“主观的”道德困境面前做出符合功利主义选择的概率约是其他人的5倍，他们更倾向于对将人推下天桥等行为表示赞同。同年，埃丽萨·亚拉梅利（elisa Ciaramelli）和朱塞佩·迪·佩雷格里诺（Giuseppe di Pellegrino）带领一组意大利学者进行研究，得出了相似的结论。意大利学者发现，健康人对于功利主义选择的抵触与更加亢奋的生理反应有关（根据手掌出汗情况进行

判断）。

近期的大量研究也都指向了同样的结论。情绪反应使人们拒绝将人推下天桥，其他“主观的”、使多数人获利的功利主义行为也受到类似的原理控制。在是否会使小火车转轨朝向家人，从而挽救5名陌生人性命的选择中，腹内侧前额叶皮层受损的病人选择肯定回答的概率比其他人更高。低焦虑性神经病患者的特点是社交情感缺陷；述情障碍则会降低患者对自身情绪状态的敏感度。这两类病人做出的道德判断都更加符合功利主义思维。面对压力时生理反应更加亢奋、本试验中表现为外周血管收缩剧烈的人，与认为自己依赖直觉的人观点相似，做出的功利主义价值判断更少。快乐是与幽默相关的正面情绪，研究人员认为快乐能够抵消参试者的部分消极情绪反应。试验表明，使人们感到快乐能够增加他们做出功利主义选择的概率。

还有一些研究指出，前文提到的杏仁核——另一个与情感相关的区域也十分重要。有精神病倾向的人思考“主观的”困境时，杏仁核的反应更加不明显。同样，阿米塔伊·沈哈夫（amitai shenhav）在我的实验室主导的研究试验表明，杏仁核的活动与类似人行天桥的困境引发的负面情绪呈现正相关，而与符合功利主义思维的选择呈现负相关。这项研究还表明，杏仁核的功能更像是最初的警铃，而腹内侧前额叶皮层则负责将这个情绪信号与其他因素进行整合，全盘考虑后做出决定。近期，莫莉·克罗克特（Molly Crockett）与同事的一项研究十分激动人心。他们让参试者服用一种类似百忧解的抗抑郁药物：西酞普兰。这种药物能在短期内加剧杏仁核和腹内侧前额叶皮层的情绪反应。研究人员让服用西酞普兰的参试者在我们设计的全套困境场景中做出选择。如他们所预测的一样，与服用安慰剂的组别相比，服用西酞普兰的参试者面对与人行天桥困境相似的、“主观的”场景时，更不易做出符合功利主义思维的判断。另一项研究表明，抗焦虑药物劳拉西泮的效用与西酞普兰刚好相反。近期，埃莉诺·阿米特（elinor amit）在我的实验室进行了一项试验，强调了视觉画面对触发情绪反应的重要作用。通过衡量人们在视觉记忆测试中的表现，可以判断出他们是否属于视觉思考者，而视觉思考占主导的参试者做出的功利主义判断更少。同样，人们进行道德判断时，对其视觉处理进行干扰则会使他们的选择更符合功利主义思维。

总之，目前已有各类证据表明，对于将人推下天桥的行为以及其他“主观的”、有害的功利主义行为，人们之所以会表示反对，是由于腹内侧前额叶皮层和杏仁体引起的情绪反应所致。然而，双加工理论的另外一面（“背侧”）又是什么呢？关于功利主义判断，我们提出了两个相互关联的假设。其一，功利主义判断来自功利主义决策标准的简单应用：做符合最多人利益的事情。*其二，面对与之相矛盾的情绪反应，人们需要额外的

认知控制才能做出功利主义判断。这里需要再次指出，在人行天桥困境中做出肯定的选择与在斯特鲁普任务中辨认词语的颜色（用绿色显示的“红色”一词）十分相似。面对相互矛盾的冲动，人类需要一个决策标准。

事实上，我们已经发现了一项能够证明这个假设的证据：人们做出功利主义判断时，大脑背外侧前额叶皮层的活动会明显加剧。而背外侧前额叶皮层恰好与自上而下贯彻规则紧密相关，也是决定人们能否成功完成斯特鲁普辨色任务的关键区域。很多其他的脑成像研究都曾得出类似的结果，但正如前文所说，脑成像数据只是相关性数据，如果我们能够对受控的认知能力施以重大影响，就像腹内侧前额叶区域的脑损伤对情绪处理造成的影响一样，那么对我们的研究将会大有益处。

对受控的认知能力施加影响的一种方式是：要求人们在做一件事的同时完成另外一项需要集中注意力的任务。我的搭档与我一同完成了这个试验，如我们所预料的一样，让人们同时完成两项任务（使人们处于“认知负荷”下）会减缓人们做出功利主义判断的速度，对于非功利主义判断的速度则没有影响。我们认为功利主义判断依靠认知控制进行，与试验结果完全一致。另一种调解认知控制的方式是给人们以时间上的压力，或是取消时间限制，鼓励人们从容做出决定。勒娜特·苏特（renata suter）与拉尔夫·赫特维希（ralph Hertwig）将这个试验付诸实践。结果是意料之中的：取消时间限制，鼓励人们从容做出决定会使功利主义判断增加。还有另外一种调解认知控制的方式，就是让人们相信，从容不迫优于迅速的直觉判断。为了营造出推崇从容不迫的气氛，可以让人们首先经历被直觉所误导的过程。乔·帕克斯顿（Joe Paxton）、利奥·昂格尔（leo Ungar）和我一同设计完成了这个试验。我们首先让参试者做一些易错的数学题目，在这些题目中，靠直觉获得的答案都是错误的答案。正如我们所料，做完这些数学题后，人们在随后进行的道德判断中更加倾向于功利主义判断。*丹·巴特尔（Dan bartels）进行的试验也呈现出一致的结果。他发现，与直觉思维相比，更喜欢费力思考的人们更容易做出功利主义判断。亚当·穆尔及其同事发现，功利主义判断与更强的认知控制能力相关。

最终，通过分析人们做出价值判断时自身所意识到的道德推理，我们能够得出很多结论。就像稍后将会解释的一样，很多因素都会在潜意识中影响道德判断。然而，在研究小火车问题的这些年当中，我遇到的所有人都能够理解将人推下天桥背后的功利主义动机。从未有人说：“为了拯救更多的生命？这是为什么呢？我从没这样想过！”赞同这种行为的原因往往是这样做的收益大于牺牲。反对这种行为的人也十分清楚，他们的选择违背了功利主义思想，但人们对此给出的理由却各不相同。解释自己的选择时，人们往往会对自己的判断感到困惑，例如：“我知道这样做不道德，但是……”阐释自己的选择时，人们也很难做到坚定不移。解释为何认为

推人的选择不正确时，人们往往会说：“这就是谋杀。”事实上，放任小火车将某人撞倒或许更应算作谋杀，但人们面临开关困境时，通常情况下都会选择赞同。总之，功利主义的原理总存在于意识当中，但人们往往对自身的反功利主义动机一无所知，这就向我们解释了情绪活动的一些重要方面。（第9章对此有更多介绍。）

## 火车轨道上的病人

哲学家们也加入了小火车困境的讨论，因为这些困境反映了一个深刻的哲学问题：个人权利与集体利益之间的矛盾。过去10年中，关于人类大脑处理类似困境时的工作机理问题，我们已经取得了很大进展。如前文所述，我们的研究甚至已经深入到分子层面。但问题在于，从这些道德果蝇身上得出的结论是否适用于现实中的道德思维？这是个好问题，却不易回答。在理想的科学世界中，我们应当设计对照试验，让人们在脑成像设备内，在认知负荷下，在腹内侧前额叶皮层受损的情况下，做出真实的、事关生死的决定，等等。但遗憾的是，这一切都是不可能实现的。退而求其次，我们面临的最好选择就是分析人们在虚拟情况下做出的事关生死的决定。

考虑到这一点，凯瑟琳·兰塞霍夫（Katherine ransohoff）、丹尼尔·温克勒（Daniel Wikler）和我共同设计了一项试验，对医生和公共卫生专家所做出的道德判断进行研究。我们向两组参试者提供了一些常见的小火车类道德困境，还提供了一些更加真实的医疗困境。例如，有些医疗困境涉及药品和设备的限量配给问题，需要就医疗资源的投入和在病人身上产生的效果进行考量。有一个场景是：是否应当将传染性病人进行隔离，以保护其他病人。另一个场景则需要在使很多人受益的、便宜的、预防性药物和只能使少数人受益的、昂贵的、治疗性药物之间进行选择。这些问题都是医疗工作中可能会遇到的真实情况。

首先，我们发现，两组参试者在传统的小火车类困境中做出的选择与他们在更加真实的医疗困境中做出的选择均高度相关。也就是说，赞同将人推下天桥的参试者更倾向于限量配给药物、隔离传染病人。这说明在小火车问题中起作用的双加工心理机制在真实世界的医疗决策中同样适用。

接下来，我们对医生和公共卫生专家可能会出现的意见分歧进行探究，这也是我们所做预测中重要的一部分。医生着眼于个别人的身体健康状况，他们有义务最大限度地避免对病人造成的主动伤害。*因此，医生可能会更加关注个人的权利。但对于公共卫生专家来说，他们面对的病人是整个社会，其首要任务是为多数人谋取福利。约翰·霍普金斯大学彭博公共卫生学院的院训便是佐证：“保护健康，挽救生命——心系大众”。因此，公共卫生专家也许会更加关注多数人的福利。这也恰好是我们在试验中得出的

结果：不论是小火车类道德困境，还是更加真实的医疗困境，公共卫生专家做出的选择与医生相比，都更加倾向于功利主义。此外，普通人与医生的选择十分类似，对功利主义的偏好程度都不及公共卫生专家。也就是说，多数人与医生一样，潜意识中都更加倾向于保护个人权利。将群体利益置于首位的思维方式似乎对人们提出了特殊的要求。

这些发现的意义在于，它们证明了道德心理学中的双加工机制不仅在实验室中成立，在现实世界也同样适用。尽管试验中的困境是虚构的，但试验中反映出的专业心态却非常真实。公共卫生领域专家在虚构的道德困境面前更倾向于功利主义选择，原因只可能是如下两点之一，或是两点的结合：其一，本身倾向功利主义思维的人更愿意进入公共卫生领域；其二，公共卫生工作者在专业培训的影响下，产生了更加符合功利主义的思维。不管怎样，这些现象在现实世界中都真实存在。面对虚构的困境，在公共卫生领域工作的人更加关注多数人的利益，我们几乎可以由此确定，这种选择与他们在真实世界的工作密切相关。同样的，如果在公共领域方面的训练会使人在实验室中的表现更加符合功利主义，那么很可能是因为相关训练会使人在工作领域变得更加功利主义。毕竟，培训的目的并非改变受训者对虚构困境的态度，而是改变他们对待工作的态度。

我们让医生和公共卫生工作者对自己的选择做出评价，结果非常发人深省。一位公共卫生专家写道："在这些极端的情境中……我感到功利主义……哲学是最适用的。从根本上来说，这才是最为道德的行为……最清晰也最公平。"与此相反，一位医生写道："如果一个人有能力决定自己的生死，并且没有主动放弃这种权利（对法律明知故犯被判处死刑），那么以他的名义决定他的生死便是对道德伦理原则的粗暴践踏。"读到这些，我似乎听到了密尔和康德的声音，听到了从地下传来的争辩。

## 两种道德思维

不论是在实验室还是在现实世界，不论是健康人还是存在情感缺陷的病人，不论是使用简单的调查问卷还是使用脑扫描技术、心理生理学以及精神药品，我们都看到了道德心理中双加工机制存在的证据。现在我们已经十分确定，人类道德思维采用的是双加工机制。但大脑为何会进化出这种机制？为什么人类采用自动的和受控的两套独立机制来处理道德问题？由于双加工机制的内在系统有时会自相矛盾，这种机制看上去问题重重。既是这样，建立统一的道德观念难道不是更加合理吗？

在第2章中，我们看到了各种不同的道德情感和下意识的倾向，从同理心到鲁莽冲动，从不负责任的欲望到对人说短道长，这些机制共同作用，促成群体内部的合作。如果这种道德观是正确的，那么人类将无辜者推下天

桥的负面反应便不过是促进合作的诸多冲动之一。（请回想第2章中库什曼的研究，他发现在实验室中模拟暴力行为会导致人的静脉血管收缩。）这些下意识的倾向已经在生物进化和文化进化过程中经受了岁月的打磨，理应为我们所用。但这些难道不够吗？为何还要建立清晰、慎重的道德思维呢？

在理想的世界中，道德直觉便已足够；但在真实的世界里，拥有双加工机制的大脑自然有其优越之处。

# 第5章 效率、灵活与大脑的双加工机制

儿子4岁那年，《虫子百科：孩子想了解的昆虫与蜘蛛的知识》这本书被我们读了好多遍。书中写道：

蜘蛛幼年时便知道如何织出完美的网。它们的行为出自本能，与生俱来。本能的优点是非常可靠，它永远能使动物以特定的方式生活。但本能的缺点是，它使动物失去了以其他方式生活的可能性。只要环境没有发生太大的变化，昆虫和蜘蛛的幼体便能够生存。但倘若它们面临新的环境，则无法通过思考摆脱困境，它们只能受本能驱使，重复自己的行为。

这段关于蜘蛛网的认知回答了“大脑为什么采用双加工机制？”这个问题，其答案是本书核心思想之一，也是过去几十年中行为科学研究提出的最重要的观点之一。

引言部分提到的比喻很好地概括了这一理论，我们之后还会反复提到。引言部分写到，人类的大脑就像是一个可以在两种模式间切换的照相机，一种是自动模式，一种是手动模式。自动模式为典型摄影场景（如“人像”、“运动”、“景物”）进行了最优设置。使用者只需按下按钮，照相机就会自动设置感光度、光圈、曝光率等数据——这就是“傻瓜”模式。这种相机还包含手动模式，使用者可以手动调整相机的各项参数。这种自动手动双模式相机为普遍存在的设计问题提供了完善的解决方案，在效率和灵活之间进行取舍。自动模式非常高效，但相对死板，手动模式则刚好相反。将两者结合起来，就可以同时享受两种模式的优点。当然，你必须知道何时需要手动调节设置，何时用“傻瓜”模式就好。

与人类不同，蜘蛛的思维中只包括自动模式，只要生活环境不变，自动模式对蜘蛛来说已经足够。但人类的生活要复杂许多，因此还需要手动模式进行额外调节。不论是个人还是群体，我们不断遇到并处理新的问题，这已是司空见惯的行为。对人类来说，只有女性别能够生育后代，但地球上几乎所有的陆地环境中，都有人类生存的足迹，这足以证明认知系统的灵活性。如果把一只丛林蜘蛛放到北极，它肯定会被冻死。但在正确的指导下，出生于亚马孙地区的孩子依然能在冰天雪地的北方生存。

人类行为的灵活性具有连锁效应。我们在发明新事物的同时，也在为后续的发明创造机会。比如，船只的发明引出了用于固定的舷外支架和用于助力的帆。我们的行为越灵活，周围环境的变化就越大，我们通过调整行为获得成功的机会也就越大。因此，在行为的灵活性方面，人类是地球上当之无愧的冠军，这也是我们能够统治地球的原因。如果有一棵树，我们可

以爬上去，也可以拿它来烧火、雕刻，可以把它卖掉换钱，也可以靠着休息，甚至可以通过数年轮来确定这棵树的年龄。具体的选择取决于我们所面对的具体机遇与挑战，我们的做法也不一定非要与我们或者他人过去的做法类似。

本章当中，我们将从相对宏观的角度探讨人类大脑的工作机制。我们将会探讨上一章提到的道德判断的双加工理论如何与双模式的人类大脑相契合。人类在生活中所有领域所取得的成功，几乎都有赖于大脑自动模式的高效和手动模式的灵活。［丹尼尔·卡尼曼（Daniel Kahneman）是支持本观点的学者中影响力最大的，他著有《思考，快与慢》（*Thinking*，*Fast and Slow*）一书，对本观点进行了精彩详细的论述。］

## 情感与理智

我们常常会把自己面临的两难处境描述为“心”和“脑”的斗争。这个比喻虽然过于简单，却能反映出人类真实的决策过程。研究人类行为的所有学科都对情感和理智的区别做出了自己的解释。但情感与理智究竟为何物？为什么人类大脑能够二者兼得？

不同种类情感的作用、起源与神经例示也不相同，因此有人认为应当将“情感”这个概念完全抛弃，但我并不认同。不同情感的共同之处并非体现在机械层面，而是体现在功能层面，事实上，“情感”与“交通工具”的概念十分类似。从机械层面来看，摩托车与割草机的原理更加相似，与帆船则相去甚远。但“交通工具”这个概念依然有其存在的意义，只不过这个意义存在于更加抽象的层面罢了。

情感是无意识的过程，你可以选择以何种方式在脑中从一数到十，但你无法选择让自己经历何种情感。你所能选择的，只是去做可能会触发相应情感的事情，比如想象你爱的人或是你恨的人。作为自发的过程，情感是为了提高行为的效率而存在的。像照相机的自动模式一样，由情感驱使的行为适用范围较广，无须刻意思考。同照相机一样，情感将环境刺激转化为相应行为结果的过程也需要过去的经验作为参考。

并非所有的无意识反应都是情感反应。大脑视觉皮层的很多烦琐工作，包括确定你看到物体的大小，将双眼所获信息统一起来等，都属于低级视觉处理过程。这一过程是无意识反应，但并非情感反应。*大脑的很多活动都是无意识的反应，比如运动时协调肌肉的收缩、调整呼吸，以及将传到鼓膜的气压波转化为有意义的信息等。事实上，大脑的大部分活动都是无意识的。那么，将某些无意识反应转化为情感反应的机理又是什么呢？

关于情感的定义，目前还没有一致的答案。但某些情感的一个重要特征

是：具有特定的行为倾向。例如，恐惧不仅是一种情感经历，还包含一系列的生理反应，让机体做好准备应对威胁：首先提高人体对当下情形的评估能力，随后使身体做好逃跑或战斗的准备。某些情感的功能可以从相对应的典型面部表情看出。恐惧对应的面部表情是眼睛睁大，鼻腔扩张，这种表情有利于人们扩大视野范围，提高嗅觉灵敏度。厌恶所对应的面部表情则刚好相反：面部紧皱可以降低病菌通过眼睛或鼻子进入人体的概率。当然，并非所有的情感都有相对应的典型面部表情，但总体说来，情感会对行为施加压力。简而言之，情感是指导人类行为的自动过程。

情感对人类行为做出的指导有时具体，有时抽象。由某件事物（比如蛇）触发的恐惧反应会对行为做出明确指导（离那东西远点！）。而其他情感状态，比如我们所说的“心情”，对于行为的影响则更加间接。这些情感状态将某些默认设置的敏感度调高，而将另一些设置的敏感度调低。最近，珍妮佛·勒纳（Jennifer lerner）和同事进行了一项经典研究，研究人员试图通过影响人们的心情，对其经济决策进行间接影响。研究人员组织部分参试者观看电影《舐犊情深》（*The Champ*）中的伤感画面，使参试者感到悲伤，结果发现情绪悲伤的参试者比其他人更愿意将自己近期购得的东西卖掉。当然，与痒感使人做出挠的行为不同，悲伤并不能直接使人卖掉自己的财物。（想象整个剧院的人眼含热泪，给各自的股票经纪人发送短信。）事实上，悲伤情绪发出的是更加开放的信号，含义类似于“事情进展很不顺利，让我们接受改变吧”。于是，当人们遇到改变的机会时，该信号就会在潜意识中使人们倾向于做出造成改变的选择。因此，有些情感提高行为效率的方式并非直接指导行为，而是对能够直接指导行为的默认设置进行调整。

理智与情感一样，都是含义模糊、真实存在的心理学现象。如果对“理智”下一个足够宽泛的定义，那么能够造成人类行为调整的一切心理过程都可归入此类。例如，你可以说，负责视觉识别的自动系统通过羽毛、喙等特征，“推理”得出：面前的物体是一只鸟。然而，对“理智”的定义也可以非常狭窄，仅限于对正式逻辑规则的有意识应用。为了叙述清晰，本书将采用一种更加中立的定义方式：理智的作用是做出决定，其含义是指对决策标准的有意识应用。斯特鲁普辨色任务能够反映出理智的一种简单形式，在这个试验中，人们会有意识地将“说出颜色”作为决策标准。在斯特鲁普任务中看到红色的单词后，我们可以将大脑的判断过程视为应用以下三段论推理的结果：“屏幕上单词的颜色是红色。我的任务是说出屏幕上单词的颜色。因此，我应当说‘红色’。”真实的推理过程也许会复杂很多，但基本的形式就是这样的。重要的是，当一个人依照理智采取行动时，他知道自己在做什么，也知道自己为何这样做；他明白自己的有效决策标

准，能够将场景中相关特点与恰当的行为匹配起来。

虽然情感的神经基质各不相同，但理智推理过程的神经基质却是大体相同的。现在我们已经知道，推理能力主要取决于背外侧前额叶皮层（DlPFC），但这并不代表推理过程仅仅发生在这一个区域。恰恰相反，背外侧前额叶皮层更像是合唱团的指挥，而不是独唱音乐家。推理过程涉及多个大脑区域，其中包括与情感反应密切相关的腹内侧前额叶皮层（VMPFC）。在理智与情感的相互关系中，有一点是不对等的：一些动物虽然有情感活动，但却没有推理（人类所谓的推理）的能力；但一切拥有推理能力的动物都有相应的情感活动。虽然有人不赞同我的观点，但我依然认为，推理本身显然是无穷无尽的。从这个角度说，就像休谟的名言所指出的那样，理智是“激情的奴隶”。（此处的激情泛指所有的情感过程，并不单指高亢的情绪。）但另一方面，理智的作用是让我们不受“激情”的影响。怎么会这样呢？

理智是情感对抗中的冠军，能使休谟所说的“平静的激情”压过“猛烈的激情”。有些价值观不能由眼前的事物自动激发产生，但理智能使我们摆脱瞬时冲动的控制，选择这些价值。但如果没有一定的情感输入（不论多么间接），理智也同样无法做出好的决定。

## 大脑的双加工过程

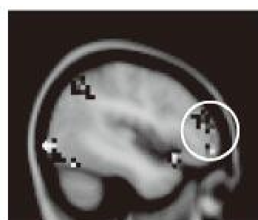
很多普通的决定都能够反映大脑的双加工结构。用随处可见的“现在”还是“一会儿”的问题来举例子。芭芭·西夫（baba shiv）和亚历山大·费德罗金（alexander Fedorikhin）设计了一项试验。试验中，他们为参试者提供了两种小吃以供选择：水果沙拉和巧克力蛋糕。对多数人来说，巧克力蛋糕是他们现在想吃的，水果沙拉则是想要留到一会儿再吃的。西夫和费德罗金将一半的参试者置于认知负荷下，让他们记忆一个7位的数字；另一半参试者的认知负荷则相对较轻，只需记忆一个2位的数字。每位参试者都需要记住数字，穿过走廊，到达另一个房间，向试验人员背出这个数字。两种小吃就放在走廊的小推车里，参试者只能取一份。试验结果是：背诵7位数字的一组，即认知负荷较重的一组，选择巧克力蛋糕的人数比背诵2位数字的一组高出50%。在事先通过问卷调查筛选出的易冲动人群中，增加认知负荷后，选择蛋糕的人数比之前翻了一倍。

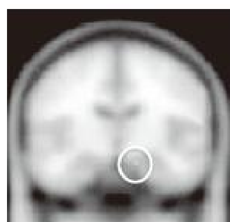
在选择小吃时，大脑中似乎有两个独立的系统在起作用。一个负责基本食欲的系统在说“给我！给我！快给我！”（自动设置），还有一个更加受控的、慎重的系统在说“别这样，卡路里太高了”（手动模式）。受控制的手动模式考虑得更加全面，既想到了现在的享受，又想到了未来的回报。但自动模式关心的只是眼下之利。如上一个章节所述，手动模式忙于其他事

情时，自动模式便更容易乘虚而入。

这种混合的食物选择方式让我们对巧克力蛋糕既想吃又不想吃。这个试验看似是认知工程学的劣等设计，但如果将其放在自然环境之中，这个设计其实非常巧妙。只要有机会，几乎所有物种在所有环境中都会优先吃掉富含卡路里的食物。在充满竞争的世界里，如果某种生物需要停下来认真思考并比较进食的效益和成本，那么它八成会错过午餐。不过，多亏有了现代科技，大部分人类享受到的午餐都是过于丰盛的。为了健康着想，也为了保持我们对他人的吸引力，我们需要保持灵活的认知，学会说“谢谢，我不吃了”。这是现代社会所特有的问题，有人应对得当，有人则稍逊一筹。但不论怎样，我们有了像样的胃口和自制力，整体的生活质量也提高了许多。当然，这并不意味着只有现代人才需要自制力：旧石器时代，饥饿的捕猎者如果不能抵挡一片熟透浆果的诱惑，便可能会错过真正的大猎物。

近年来，认知神经科学家对人类的“推迟满足感”进行研究，在神经的诸多功能中，发现了另一种现在已经广为人知的神经特点。山姆·麦克卢尔（sam McClure）和他的同事在一次研究中要求参试者面对两种类型的选择，分别做出决定。一类选择中包含可以立即获得的奖励*：现在获得2美元，还是下周获得3美元？另一类选择中只含有稍后才能获得的奖励：下周获得3美元，还是下下周获得4美元？研究人员发现，立即获得奖励的期望激发了包括腹内侧前额叶皮层在内多个大脑区域的活动，但背外侧前额叶皮层在做出所有决策的过程中都保持着活跃。此外，人们选择立即获得奖励（给我！快给我！）时，第一组显示的大脑区域会更加活跃，而选择稍后获得更多奖励（“想想以后……”）时，第二组中显示的大脑区域则会更加活跃（参见图5.1，第一行）。



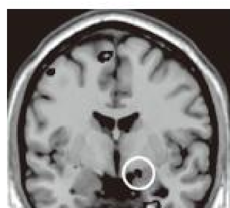




图5.1 三个脑成像试验的结果揭示了自动情感反应（“自动模式”）与受控认知（“手动模式”）之间的相互影响

我们要注意到，在某种程度上，选择稍后获得更多的奖励就好像选择将人推到火车轨道上一样。因为在这两个场景中，我们都需要背外侧前额叶皮层来抵御腹内侧前额叶皮层所控制的情感倾向，选择“多数的利益”。当然，两种情况之间也有着重大的差异。“现在”还是“以后”的选择中，情感信号反映的是个人欲望；人行天桥困境反映的则是对他人的道德考虑。同样，在“现在”还是“以后”的选择中，我们考虑的是个人利益的最大化（个人的）；人行天桥困境中，我们考虑的是多数人的利益最大化（人与人之间的）。但不管怎样，在最为抽象的功能层面，以及功能神经解剖学层面，我们看到了相同的工作机理。

人们试图调节自身情绪时，我们也发现了相似的机理。在试验中，凯文·奥克斯纳（Kevin ochsner）及其同事将易于引发强烈负面情绪的图片（如在教堂门口哭泣的妇女）展示给参试者，然后要求参试者以更加积极的方式重新诠释图片内容。比如，想象哭泣的妇女并非悲伤的悼亡人，而是在婚

礼上喜极而泣的宾客。如果仅仅观看这些图片，大脑中活跃的区域依然是与情绪相关的老朋友——腹内侧前额叶皮层和杏仁核。然而，在重新理解图片的过程中，背外侧前额叶皮层也变得活跃起来（参见图5.1，第二行）。此外，由背外侧前额叶皮层控制的重新阐释过程对杏仁核和腹内侧前额叶皮层的活动产生了抑制作用。

事实上，大部分人遇到其他种族的群体外成员时，似乎都会自发地进行类似的重新考量。维尔·坎宁安（Wil Cunningham）及其同事在试验中向白人参试者展示了许多白人和黑人的头像。有的头像以潜意识的方式出现，展示时间仅有30毫秒，参试者几乎无法进行有意识的感知。有的头像可能会保持0.5秒，使参试者能够进行有意识的感知。以潜意识方式出现的头像中，黑人头像比白人头像更能激发白人参试者大脑中杏仁核的活动（参见图5.1，第三行）。此外，对于在内隐联想试验中对黑人做出更多负面联系的参试者来说，这种差别会更加明显。

本次试验中，研究人员鼓励所有参试者在观看头像时放弃种族偏见，所有参试者的努力都能在脑扫描结果上反映出来。只要屏幕上的头像停留足够长的时间，使参试者能够有意识地感知，背外侧前额叶皮层的活动便会增加（参见图5.1，最后一行的右侧），杏仁核的活动会减弱，与奥克斯纳的情绪调节试验结果一致。一个后续试验也同样证明，对于不愿成为种族主义者的白人参试者来说，与黑人互动会使认知负荷增加，导致参试者在斯特鲁普辨色任务中表现不佳。

因此，大脑的双加工机制不仅在道德判断过程中有所体现，而且还体现在食物、金钱的选择以及我们想要改变的态度等其他方面。大部分情况下，大脑的自动设置会指导我们如何行为，但我们也可以切换到手动模式，更改默认设置。但前提是我们对自己切换模式的能力有所了解，并且愿意进行模式切换。

## 变聪明些

到现在为止，根据我的描述，你可能认为大脑的自动模式只会惹麻烦：使我们变胖、沮丧、变成种族主义者等。但事实上，这些有害的冲动只不过是例外情况，大脑的自动模式通常情况下是非常智能的（参见第2章）。保罗·惠伦（Paul Whalen）及其同事的试验显示，人类能够极其迅速地识别表示恐惧的面部表情，只要在眼前停留1.7%秒，杏仁核便能够做出反应。如此迅速的反应，自然有其诀窍：大脑并不会对整个头像进行详细分析，而是只对恐惧情绪的面部特征进行捕捉，即放大的眼白（参见图5.2）。



图5.2 左侧：恐惧情绪会促使杏仁核扩大。右侧：儿童读物《晚安，大猩猩》中，动物园管理员的妻子发现，好几只大型非洲动物溜进了自己的卧室

腹内侧前额叶皮层也同样非常智能。如达马西奥的试验团队发现，腹内侧前额叶皮层可以帮助人们评估风险，做出决策。在一个经典试验中，参试者需要从4堆牌中随机抽取，每抽一张牌后，参试者都会相应地赢得或者输掉一些钱。4堆牌中有两堆好牌，即总体上讲，这两堆的牌能使参试者赢钱；另外两堆牌则是坏牌，即如果赢牌，赢得的金额会变大，但如果输牌，则输掉的金额也会更大，总体来说，这两堆牌会使参试者输钱。起初，参试者不知道哪堆是好牌，哪堆是坏牌，只能随机抽牌，看看自己的运气。健康的参试者很快就能针对不好的牌堆形成负面反应：他们伸手去不好的牌堆抽牌时，手掌会出汗。*更奇妙的是，早在人们辨认出不好的牌堆之前，他们伸向坏牌堆的手就已经开始出汗了。然而，腹内侧前额叶皮层受损的病人却无法感知这样的生理信号，他们会一直不断地从坏牌堆抽牌。也就是说，健康人大脑的腹内侧前额叶皮层能将根据经验获得的多条信息进行整合，将这些信息转化为情感信号，向决策者提出中肯的建议，从不同牌堆抽牌的经验总结便是如此。此外，早在人们能够感知好坏及其原因之前，这种第六感式的建议大概就已经形成。这就解释了之前提到的腹内侧前额叶皮层受损病人的问题：虽然他们在标准实验室推理测验中表现良好，但在生活中却常常做出糟糕的决策。他们“知道”，但不能“感觉”，而感觉恰恰是非常有用的一个环节。

因此，人类情感中的自动设置与手动设置都十分有用，我们在不同的情况下需要不同的模式。摄影时，自动模式在相机设计者已经预想到的场景中

十分有用。例如，在柔和的室内光线下给5英尺外的一个人拍照（“人像”），或是在明亮的阳光下为远处的山峦拍照（“景物”）。同样，对于大脑来说，基于过去的经验“设计”出的自动模式也是最好用的。

类似的经验有三种形式，分别基于三种不同的试验和错误得出。其一，由基因决定。尽管早已逝去的祖先们不再有机会将他们的基因传承下来，大脑会将他们来之不易的经验进行整合。其二，通过文化学习获得。先人们不断探索，不断失败，他们的想法对我们产生重大影响。多亏了他们，我们不必亲自遇到纳粹和三K党就能发自心底——确切地说，是发自杏仁核——地知道，纳粹的标识和头戴白色尖顶兜帽的人带来的都是糟糕的消息。其三，老生常谈的个人经验。孩子通过痛苦的体验明白，滚烫的火炉不能摸。我们的“直觉”未必像蜘蛛一样与生俱来，但如果直觉有用，它一定是来自某个人的经验，这个人可以是你自己，可以是生物学层面上的祖先，也可以是文化层面上的“祖先”。

人类大脑的手动模式，即认知控制能力，与自动模式的工作原理有着本质不同。事实上，认知控制专门解决自动模式所无法解决的问题，可以用学习开车进行说明。很明显，人类没能继承会开车的基因，渴望开车的年轻人会沮丧地发现，在文化层面熟悉开车并不能使他们学会开车，或者说学会安全地开车。当然，驾车新手无法依赖个人经验，因为个人经验恰好是他最缺乏的东西。事实上，学习开车需要极大地调动背外侧前额叶皮层，如果你第一次坐上驾驶位就想依靠自动驾驶仪，那么最后你也许只会绕着大树打转。

所以，想要变聪明有三个条件。第一，从生物学祖先、从周围的人，甚至从自己的经验中学习能够广泛应用的本能。第二，变得聪明需要使用手动模式，需要谨慎处理复杂及异常问题的能力。第三，我们需要元认知技巧，就像摄影师的摄影技巧一样。与照相机不同，没人能告诉我们何时使用“傻瓜”模式，何时使用手动模式，我们需要自己做出决定。不论出于个人考虑还是作为新草地上努力共处的牧民，如果我们能够理解思维的工作机理，也许就能做出更加明智的决定。

# 第三部分 通用货币慢

# 第6章 绝妙的想法

作为新草地上的牧民，我们应当如何化解分歧？如何避免常识道德悲剧？这就是我们试图解决的问题，现在的我们已经具备了寻找解决方案的条件。让我们先来回顾一下。

第1章中，我们将常识道德悲剧与最初的公地道德悲剧进行对比。公地道德悲剧中，自私是合作最大的障碍，道德则是大自然提出的解决方案：道德能使人类在“我”和“我们”之间优先考虑“我们”。但如何让“我们”与“他们”和谐共处，如何解决常识道德悲剧，大自然并没有给出现成的答案，这也正是我们所要思考的问题。要想避免常识道德悲剧，我们就需要自己找出一种非自然的解决方式：我将这个答案称为“元道德”，是一种更高层面的道德体系。某个部落的道德观可以裁决个人利益冲突，而元道德则用于裁决不同部落间的道德观冲突。

第2章中，我们对大脑道德机制的标准状况进行了审视。幸运的是，人类大脑本身便配有自动的行为机制，能够促进并维护人际合作和群体间合作。这些机制包括同理心、复仇、荣誉感、内疚、羞愧、部族主义、义愤等。这些社会冲动能够平衡人类自私的冲动，使我们进入神奇角落，避免公地悲剧。

第3章中，我们再次将目光投向新草地及我们所引进的道德机制。人类的道德思维对于促进群体内部合作（“我”与“我们”）十分有用，但解决群体间合作问题（“我们”与“他们”）则并不十分有效。从生物学角度来看，这是情理之中的状况。因为生物学认为，人类大脑的基本机制是促进群体内部合作，*鼓励群体之间竞争。部族主义（群体层面的自私）、合作术语方面的分歧（个人主义还是集体主义）、对本地文化“专有名词”的维护（领袖、神灵以及圣书等）、有偏好的公平，以及带有偏好的事实观念等因素都阻挠了群体间合作。

第一部分的三个章节中，我们将人类大脑描述为一系列自发冲动的集合。自私的冲动使社会生活充满挑战，而道德的冲动则使社会生活成为可能。第二部分中，我们对人类大脑有了进一步的了解，就像可以在两种模式间切换的照相机一样，大脑也包括自动和手动两种模式。自动模式下的情绪反应传承从前的基因，从文化经验和个人体验中总结出成套的经验，让我们迅速有效地做出决定。手动模式则是一种有意识的、外显式的、实用的推理能力，让我们做出灵活的决策。快速思考和从容思考之间的矛盾在道德困境中尤为突出，比如人行天桥困境中，直觉会说，“别把那个人推下天桥！”但有意识的、基于原则的道德推理则会说，“但这样能挽救更多生

命！”正如上一章所述，直觉与推理之间的矛盾并不仅限于道德问题，而是普遍存在于人类大脑中，即使是吃巧克力蛋糕还是吃水果沙拉这样的日常选择，也体现了同样的矛盾。

新草地问题的解决方案就包含在这些概念之中吗？答案是肯定的。第一部分和第二部分各自阐释了一种解决方案，一种是哲学层面的，另一种是心理学层面的。但事实上，这两种解决方案是殊途同归的同一方案。下面我们便从哲学层面的方案开始讲起。

## 绝妙的想法

崇尚个人主义的北方牧民认为，好的牧民应当为自己的行为负责，遵守自己的承诺，尊重他人的财产，仅此而已。崇尚集体主义的南方牧民认为，好牧民的特质不止于此，在他们看来，公正的社会中，生活的重担与甜蜜应当平均分配。不同部落间还存在许多其他的分歧，比如，关于荣誉、关于谁先动手、谁下手更狠、谁的话语绝对正确、谁值得我们宣誓效忠、谁应当获得原谅，以及在全能的上帝面前，何种行为无法获得宽恕等。既然各个部落关于道德生活的观念各不相容，那么他们在新草地上应当如何相处？

解决方案之一是：没有正确答案。有些部落这样做，有些部落那样做，大家就这样各行其是。这就是众所周知的“道德相对主义者”给出的答案。* 这个答案的问题在于：这几乎不能算作一种方案。在一些重大问题上，道德相对主义者也许是正确的，正如他们所说，终极道德真理也许并不存在。然而，即使他们是正确的，人们注定要按照各自的方式生活，道德相对主义者不愿从中做出选择，但总有人不得不进行选择。道德相对主义者可以拒绝选择，但这本身其实也是一种选择，它代表了对道德判断的一种评判态度。即使道德真理真的不存在，但道德选择却是无法回避的。

如果道德相对主义不是避风港湾，那我们该怎么办？我们自然会想：也许牧民们应当简单地选择效果最好的做法。在新草地上，如果个人主义比集体主义有效，就采取个人主义；如果集体主义更加有效，就选择集体主义；如果严格的荣誉准则能维护和平，那我们就培养荣誉文化；如果荣誉文化导致争斗不断，那么我们就摒弃它，依此类推。

在我看来，这是一个绝妙的想法，本书余下的部分将不遗余力地发展这一想法，并为其辩护。相信你已经注意到，这是符合功利主义思维的想法，从抽象层面来说，这是符合结果主义思维的想法。（后文会针对这一点详细论述。）理论上来讲，选择最有效的做法对很多人都是理所当然的：谁不愿选择最有效的做法呢？但正如第4章阐释的那样，在具体的道德问题面前，我们是否应当选择导致最好结果的做法？答案似乎不再显而易见。

即使我们假设将人推下天桥能够在相对情况下得出最好的结果，但这种做法对很多人来说似乎依然不对。此外，在社会的组织运转方式上，这种计算成本效益的思维方式与很多人心底的价值观也是相互矛盾的。为了便于解释，我们依然从可爱的故事人物开始讲述。

## 长老的智慧

对于选择最有效做法这一观点，如果你向北方牧民询问意见，几乎所有人都会表示赞同。他们会声称自己偏爱最为有效的体制，当然，个人主义就是最有效的体制。如果你向南方牧民询问意见，他们的回答与北方牧民相类似，却又相互矛盾。崇尚集体主义的南方牧民会说，在新草地上生活的牧民应当遵从最为有效的体制，毫无疑问，集体主义就是最有效的体制。

这是怎么回事？也许，北方牧民和南方牧民的道德观在根本上是一致的：他们都希望选择最有效的体制。两方的分歧仅仅是事实认定上的分歧：究竟哪种体制最为有效。为了验证这个想法，可以做一个思维试验。假设我们给崇尚个人主义的北方牧民拿出堆得像小山一样的证据，表明集体主义更有效；也给崇尚集体主义的南方牧民拿出堆得像小山一样的证据，证明个人主义更加有效。两个部落都无法对这些证据提出质疑，因为他们对于另外一种社会状态几乎没有任何概念。他们将如何应对这个挑战？也许有些北方牧民的兴趣会被激发起来，但对大多数人来说，这些所谓的证据不过是一派胡言，不足让人深究。（请回想第3章中我们对种族偏见的探讨。）大多数的南方牧民也会做出同样的反应。如果是这样，那么两个部落的分歧似乎与事实认定并无关系。

还有另外一种验证方式，这次我们不再向北方牧民和南方牧民展示证据，支持另外一方的生活方式，而是要求双方自己想象这样的证据。对于崇尚个人主义的北方人，我们假设：在集体主义至上的社会里，人们生活得更好。因为在个人主义至上的社会中，总有赢家和输家之分，有些人牲畜成群，有些人却几乎一无所有。但在集体主义至上的社会里，并无输赢之分，每人的财产虽然不多，但都很平均。个人主义至上的社会中财富总量更大，但总体来看，社会状况却要更差，因为在这样的社会中，输家失去的利益远远超过了赢家获得的利好。但在集体主义至上的社会中，情况则正相反，没人富得流油，但大家都能丰衣足食，整体社会状况更好一些。我们询问北方牧民：“如果上述情况都是真实的，你们会转而投向集体主义吗？”

首先，我们的北方朋友会说，这个问题实在很蠢，因为所有人都知道，集体主义意味着灭亡之路。然后，我们的北方朋友会举出一系列例子，讲述集体主义经过不同的发展路径，最后终将走向灭亡的过程。他们还会说，

集体主义者是只想占人便宜的懒人，或是不能正确认识世界的幼稚之人，抑或是与集体主义者在一起太久，思想被同化的人。我们会礼貌地点头，然后提醒他们，我们问的并不是真实世界里的集体主义如何发展，而是一个简单的假设：如果集体主义真的更加有效，你们会不会改变观点？这样的阐述和解释反复进行几轮之后，终于有一些北方朋友愿意对这个冒昧的问题做出回答。他们高昂着脑袋，回答说如果在某个荒诞不经、黑白颠倒的世界里，集体主义真的更加有效，那么向集体主义倾斜也是可以考虑的选择。

接下来，最为睿智的北方牧民，即北方的长老，会走上前来，向我们解释集体主义不仅在实际生活中会造成毁灭性的后果，其本身的哲学内涵也早已从内部开始腐烂。他会说，如果笨人和懒人因为笨拙和懒惰不能为自己赢得财富，那么他们也不应要求分得财富。社会的进步与否不应以其分配给所有公民的福利总量进行衡量，社会进步应体现在对于正义的追求。北方长老认为，集体主义是不公正的，它惩罚了最努力的人，却奖励了最懒惰的人。话音刚落，北方牧民人群中便爆发出一阵掌声，他们的核心价值观在这番演说中得到了最好的体现。

对于崇尚集体主义的南方牧民，我们提出了相反的假设：如果最终发现个人主义更加有效，你会转而倾向个人主义吗？同北方的同伴一样，南方牧民最初也对这个假设不屑一顾。他们说，人人都知道，如果一个社会以个人主义的贪婪作为基础，那么这个社会注定走向毁灭。我们同样重申，这只是一个假设的问题。也同北方牧民一样，几位南方牧民犹豫不定地表示，如果个人主义真能以某种方式变得更加有效，那么也是值得考虑的。然而，南方长老带着长者的智慧和权威对这种妥协的态度进行了批判。她解释道，个人主义不仅在实际生活中会引发诸多悲剧，其本身的哲学内涵也早已从内部开始腐烂。一个以贪婪作为基本准则的社会，其本身就是不道德的。南方的长老认为，高贵的牧民关于爱、怜悯和情义的理想千金难买。话音刚落，南方牧民人群中齐声唱诵：阿门。

这些都是思维试验，我自然有权想象其结果。但我想象出来的结果事实上是根据我们学过的道德心理学严格推理出来的。北方牧民崇尚个人主义，并不是因为他们认为个人主义更加有效，他们选择社会制度时，并没有进行成本效益分析比较。同样，南方牧民的集体主义情结也并非因分析而产生。事实上，北方牧民和南方牧民之所以秉持自己的信念，是因为他们终生都在各自部落文化的浸润下生活，其道德本能也与各自的生活方式相适应，在各自的社会系统中避免公地悲剧的发生。如果不发生戏剧性的文化转变，不论事实如何，集体主义在北方牧民眼中永远都是错误的；个人主义在南方牧民眼中也永远都是错的。双方都坚信其各自的生活方式更加有效。但最关键的是，与选择更有效的社会制度相比，双方最看重的却是保

持各自的生活方式。长老们深知这一点，他们是地方智慧的守护者，不会因我提出的假设而上当。他们深知，从根本上说，己方的价值观与“何者更加有效”没有关系，他们赖以生存的是更深层次的道德真理。

## 结果主义、功利主义和实用主义

从前，人们认为寻找最有效做法是一个绝妙的选择，但他们发现，自己真正想要的有时候并不是最为有效的。但不管怎样，最有效的做法依然是上佳的选择，这就是功利主义思维背后的依据。这是十分现代的哲学思想，人们却常把它当作简单的生活常识。

“选择最有效的做法”这个观点听上去像是通俗的“实用主义”。（美国哲学思维中，“实用主义”通常指代另一种含义。*）但功利主义的含义远不只是讲求实用。首先，“实用主义”可能包含对眼下权宜之计的选择和对长远利益的放弃，但这并不是我们心中所想。功利主义认为，我们所做的应当是真的最有效的，不仅是为了眼下，即使从长期来看也应如此。其次，“实用主义”所指的可能只是灵活的处事风格，适用于任意一种价值观。从通俗意义上讲，坚定的个人主义者和坚定的集体主义者都可以是“实用主义者”。但与此相反，功利主义是一种核心价值观，需要将“实用主义”贯彻到底，不论结果如何，都将其作为基本原则来执行，即使这种做法与部族本能相违背也不能放弃。

因此，我更愿意将功利主义视为深度实用主义。还有其他原因支持我的观点，稍后会对此进行介绍。如果与你约会的人说：“我是一位功利主义者。”也许你还需要更多时间对他/她进行观察。但如果他/她自称是“深度实用主义者”，那么也许你就可以带他/她回家，介绍给父母了。“功利主义”这个词听上去并不顺耳，而且容易造成误会，因此我们最好抛弃这个名字。然而作为一名深度实用主义者，我也深知自己不可能仅凭几句话就改掉一个已有200年历史的哲学术语。我需要证明，深度实用主义才是我们所寻找并想要的结果，这并非旧瓶装新酒，换汤不换药的做法。因此在眼下，我会暂时向传统致敬，使用这个不顺耳、容易造成误会的传统名称来指代我们的绝妙想法。在第五部分，当我们充分理解功利主义，了解如何应用这种思维时，我们再回过头来重新说明，功利主义事实上就是深度实用主义。

那么，什么是功利主义？这个概念究竟从何而来？首先，功利主义是结果主义的一种形式。迄今为止，我对功利主义所有的描述都适用于范围更广的结果主义。结果主义认为，“结果”（即实用主义者所说的“效果”）是衡量一切的终极标准。请注意“终极”这个重要的词，也就是说，除结果以

外，其他因素（比如：诚实）并非不重要，只是结果应当作为标准，衡量其他因素是否重要。依照结果主义的观点，我们的最终目标就是使事情尽可能顺利。*

但这里“顺利”的含义是什么？应当据何判断结果的优劣？功利主义对此问题给出了具体的回答，这也是功利主义与结果主义之间的区别之一。结果主义的观点听上去好像在做“成本效益分析”，从某种程度上说，也确实如此。但人们往往会将“成本效益分析”与金钱联系起来。作为牧民，我们或许应当以经济生产率作为衡量成功的标准：能够最大限度地提高GPP（草场生产总值）的机制便是最有效的机制。这样一来，道德的计算便得以简化，因为物质财富总是更加便于衡量比较。但是，经济生产率真的是最重要的因素吗？我们可以轻而易举地构想出一个社会，其经济高度发达，人民却水深火热。这样的社会是好的吗？

如果经济成果不是真正重要的结果，那么真正重要的是什么呢？我们可以扪心自问，我们希望从经济生产率中得到什么？再次重申，如果我们都生活在水深火热中，财富显然没有任何帮助。反过来说，如果我们都生活得快乐幸福，贫穷或富有也许并不重要。因此我们会自然想到，幸福才是最重要的因素。并非所有人都赞同这一点，但至少这个观点相对合理。将“幸福是最重要的因素”和“我们应当尽量扩大好的结果”两个观点相结合，就得到了功利主义的观点。

作为功利主义的奠基人，边沁和密尔并不是空想哲学家，相反，两人都是大胆的社会改革家，积极致力于解决当时的社会问题和政治问题。事实上，我们现在所熟知的很多社会问题之所以会成为社会问题，还要归功于边沁和密尔两人。两人的观点在当时十分激进，但现在，他们为之奋斗的很多社会改革在我们看来已经是理所当然。在反对奴隶制、倡导言论自由、市场自由竞争、普及教育、环境保护、监狱改革、女性权利、动物权利、同性恋权利、工人权利、离婚自由、政教分离等方面，两人都是时代的先行者。

边沁和密尔都不愿因惯例而接受道德规则，他们不会因为某种行为或政策符合传统，或因为大多数人对其感觉正确，或因为那是“事物的自然秩序”就对其表示认可。他们也并未搬出上帝来证明自己的道德观点。事实上，他们提出的问题就是我们在前文提到的问题：真正重要的是什么？为什么它是真正重要的？如果不回避问题的实质，我们应当以何种标准衡量自己的行为和政策？一个人基于何种观点才能够认定奴隶制是错误的？边沁和密尔无法请出上帝为自身辩护，因为反对奴隶制的声音认为上帝站在他们一边。退一步说，即使功利主义的两位创始人想要借上帝为自己辩

护，他们又如何证明自己对于上帝意愿的解释是正确的呢？出于相似的原因，两人也无法利用奴隶的权利为自己辩护，因为奴隶的权利问题也是具有争议的话题。既然如此，人们应当根据什么来判定何人拥有何种权利呢？

边沁和密尔从功利主义的角度给出了答案。衡量法律和社会实践时，他们只考虑一个问题：这种做法会增加还是会减少人们的幸福感？程度有多大？比如，他们认为奴隶制是错误的，并不是因为上帝反对奴隶制，而是因为奴隶制带来的好处（例如，提高经济生产率）远比不上它造成的悲剧。同样的道理也能解释限制妇女自由、野蛮对待动物、立法禁止离婚等做法的不合理性。

为了衡量道德观，帮助人们在艰难的道德抉择面前做出决定，边沁和密尔引入了一项普遍适用的标准：以所有行为对人们幸福感影响的总和作为衡量标准。功利主义是绝妙的想法，我相信这就是这些现代牧民迫切需要的元道德。但与此同时，功利主义也是备受争议的想法，哲学界对此已经进行了长达两个世纪的争论：所有的道德观都能用简单的标准进行衡量吗？如果答案是肯定的，那么幸福感的总和能够作为正确的标准吗？在后面的章节，特别是第四部分中，我们会对功利主义面临的哲学问题进行讨论。但首先我们要解释清楚，什么是功利主义？为什么有人认为幸福是最重要且唯一重要的因素？

## 功利主义的理解和误解

正如第4章所述，人们对功利主义普遍存有误解，所有麻烦都始于这个糟糕的名字，听上去过分强调了世俗的功能性。（“功利的房间”指的是洗衣服的地方。）用“幸福”代替“功利”，我们便向着正确的方向迈出了第一步。但这一步也同样具有误导性，因为功利主义哲学家所说的“幸福”比我们心中“幸福”的含义宽泛许多。即使我们能正确理解“幸福”的含义，“将幸福感最大化”这个想法也极易引起误解。有人认为，功利主义者的生活中充满了计算，做出每个决定都需要计算得失，但事实并非如此。事实上，“将幸福感最大化”这项任务显得含糊不清，让人摸不着头脑：幸福的含义是什么？不同人眼中的幸福难道不是各不相同吗？如何衡量幸福？什么算作幸福，如何增强幸福感，这两个问题由谁来决定？“将幸福感最大化”难道不是危险的乌托邦主义吗？下面的内容中，我将会解答这些问题，对常见的误解做出回应，为这个臭名昭著的哲学思想正名。

### “幸福”的含义是什么？

中学时，我和班里的同学一起完成了一个“价值观项目”。我们每位同学都

写出自己认为最重要的10项价值观，并说明它们为何如此重要。我们把汇总的结果编成了一本书——事实上，那是一本满是图画、照片和手写注释的剪贴簿。各自开始之前，我们在一起做了头脑风暴，同学们提出不同的价值观念，由老师把它们写到黑板上，包括“家庭”、“朋友”、“宗教”、“运动”、“开心地玩耍”、“爱”、“帮助他人”、“学习新知识”、“我的猫咪”。然后，我们对一些概念进行整合：把猫咪归入“宠物”，把“迪士尼乐园”归入“开心地玩耍”等。于是我们得到了一张不错的清单，“幸福”这个概念也赫然在列。

作为一名新晋功利主义哲学家，年少的我不知该如何将“幸福”列入清单。首先，我写下了“家庭”，随后是“朋友”，“幸福”大概排在第四位。但我总会想到，清单上的所有事件其实都与幸福相关：如果我认为家庭和朋友非常重要，那么我所看重的，不就是他们自身的幸福和他们给我带来的幸福吗？如果“运动”和“爱”不能带来幸福，还会有人看重它们吗？因此我想，这些概念还需要进一步梳理。

显然，我的同学并没有这样的担忧。在他们看来，幸福不过是清单上的一个事件而已，这一点没有任何问题。他们为什么会这样认为呢？也许，他们心中所想的，与大多数人提及幸福时心中所想的一样：看重幸福的含义就是看重那些能让人微笑的事情。电影《音乐之声》（*The Sound of Music*）中有一首曲子叫作“我最喜爱的东西”，歌中唱出了很多让人微笑的事情。

在音乐剧《你是个好人，查理·布朗》（*You’re A Good Man*， *Charlie Brown*）中，“幸福是什么”这支曲子给出了进一步的解释：

幸福是两种冰激凌

一种是知晓一个秘密

一种是爬上一棵树

现代成年人心中，有一张诱人的生活清单，幸福就是单子上列出的各种闲适活动：倚在轻轻摇动的吊床上，用iPad（平板电脑）阅读新闻；和邻居闲谈几句，然后出门去山地骑车度过一个下午；日落时分在甲板上小酌一杯……我们把这类幸福的概念称为“最喜欢的事情”。如果幸福指的是沉浸在最喜爱的事情中尽情享受，那么它确实只是清单上的一项普通事件。此外，如果幸福就是做自己最喜爱的事情，尽情享受，那么将其视为终极价值，作为衡量一切行为的标准这个想法也似乎过于浅薄。

然而，把幸福定义为“最喜欢的事情”这个概念经不起推敲。问题在于，我们谈及幸福时，往往不会想到可能对幸福感产生巨大影响的因素。比如，更换汽车刹车片并不是我喜欢做的事，但如果不换刹车片，很多人（我自己，我的家人，其他开车的人，他们的家人）的幸福都会因此受损。再比如，一个人奋斗多年，投身于一项意义重大但困难重重的事业，在我们列出的价值观清单中，这个人的行为似乎更应归于“努力”、“坚持不懈”、“自律”，而不是“幸福”。但毫无疑问，这项事业的建立最初一定是为了增进某个人的幸福，即使不是为了奋斗者本身的幸福，也一定是为了某位他人的幸福。同样的，一个人之所以会去流浪汉之家做义工，并不是因为她享受这个过程，而是因为在她看来，对更加不幸的人伸出援手十分重要。这项行为所体现的价值观更像是“帮助他人”、“慈善”和“社会责任”，而非幸福。但我们又可以想到，志愿者的愿望是改善穷人的生活质量，而生活质量的改善则包括自身幸福感的提升和使他人感到幸福的能力。这个例子说明，将幸福定义为“最喜欢的事情”还有另外一点不妥：这个概念忽视了幸福这把标尺上数值为负的部分。如果人类的最终目标是将幸福感最大化，那么帮助人们摆脱贫困无疑比冰激凌顶端的樱桃更加重要。但是人们谈及幸福时，浮现在脑海中的画面更多的是冰激凌顶端的樱桃，而不是流浪汉之家。

因此，幸福是深植于很多价值观内部的，而且乍看之下，这些价值观都比幸福显得更加深刻、更加有意义。上文提到了家庭、朋友和爱，还包括知识、真理、教育、艺术等智力价值；自由、正义等公民价值；勇敢、诚实、善于创新等性格特质等。所有这些都让世界变得更加幸福，我们之所以会看中这些价值，也是出于这方面的考虑。概括来讲，几乎所有值得重视的价值观都与幸福紧密相关。这至少说明，幸福作为一种道德价值，很容易被误解和低估。要想更加准确理解幸福的含义，我们需要抽象思维，具体来说，就是反事实思维。站在功利主义的角度来看，看重幸福并不仅是看重通常情况下由幸福而联想到的事件。只要某件事情的缺失会造成幸福感降低，这件事情就应当受到重视。如此说来，我们重视的所有事情几乎都被囊括在内了。

然而，功利主义的观点不仅限于指出幸福的重要性，而是指出幸福（含义完整的幸福）是唯一重要的事情。这是为什么呢？要想解释原因，我们需要从眼下最紧迫的关切开始，然后向前倒推，不断询问自己，“我为何在意这一点？”直到自己再也无法回答为止。比如，今天你去上班了，这是为什么呢？也许因为你很享受工作，也是为了挣钱。那么你为什么需要挣钱呢？为了买食物。你为什么需要食物？因为你和家人喜欢吃东西，不喜欢挨饿，并且食物能让你和家人活下去。但你和你爱的人为什么愿意活下去呢？因为你们享受活着的感觉，尤其是生活中相依相伴的感觉。为什么你会在意自己和家人是否享受生活呢？呃……

将这种逻辑重复多次，就会发现，你做的所有事情的目的都是提升个人体验。即使是惩罚等不愉快的事情也是如此。虽然惩罚的直接目的是产生不愉快的体验，但它仍与提升体验有关。施以惩罚让人感觉不快，从而阻止坏事的发生。这就间接提升了潜在受害者的生活体验。通常来说，石头等没有体验能力的事物并不在道德考虑范畴之内。

因此，考虑到个人体验的质量，一切事物的优点和缺点似乎最终都会消失。如此看来，值得重视的价值有很多：家庭、教育、自由、勇敢，以及黑板上列出的所有价值观。但在功利主义者看来，这些价值之所以重要，只是因为它们会对人的体验产生影响，如果这些价值无法再对人类体验产生积极影响，它们的价值也就不复存在了。总之，如果某件事不能影响人类体验，那么它便是不重要的。

这就是功利主义定义下的幸福背后的逻辑。幸福不仅是美味的冰激凌，是湖边小屋外的暖夏傍晚，更是个人所获体验的整体质量。重视幸福就意味着重视一切能够提升任何人体验质量的事情，对于生活体验亟待提高的人们来说尤为如此。对于功利主义者来说，幸福并不一定优于清单上其他的价值观。事实上，如果能够正确理解其含义，幸福与其他价值呈现包含关系，幸福就像是标准模型中的希格斯玻色子。幸福是为其他价值观赋予价值的价值观。

对于这个观点，你或许同意，或许反对，后文将会提到，我认为这个观点有些夸大其词。（如果幸福真的是唯一的终极价值，那么不同价值之间为何还会存在冲突？）目前为止，我想要表达的意思有两层：第一，功利主义定义下的幸福含义十分广泛，包含了所有积极的体验，排除了所有消极的体验，这就是我们所说的“幸福”的含义。第二，基于这个观点，便可以理解，我们为何认为幸福与价值观清单上的其他项目不同，在价值观体系中占据特殊位置。与多数的激进功利主义者不同，我并不认为幸福是唯一正确的价值观。在我看来，边沁和密尔思想的精髓是：幸福之所以地位特殊，是因为幸福是人类价值体系中的通用货币。接下来的两章中，我们还会继续讨论这个观点。

## 是否有些幸福比其他幸福更重要？

前文提到，功利主义定义下的幸福概念十分宽泛。但事实上，关于这一概念的宽泛程度，功利主义者之间也存在争议。边沁的概念中，功用的概念十分狭窄，只包括快乐和痛苦两个方面。密尔的概念则相对宽泛，他认为有些快乐从本质上讲便优于其他快乐，也因此比其他快乐更加有价值。密尔有一段名言：

宁做不快乐的人，不做快乐的猪；宁做不快乐的苏格拉底，不做快乐的傻

瓜。如果傻瓜和猪做出的决定与此相反，那是因为他们只能看到问题的一面，而苏格拉底和人却能看到完整的两面。

一方面，将福祉的定义仅限于快乐和痛苦似乎并不明智，对所有的快乐一视同仁也不是聪明之举；另一方面，密尔对某些更加“高级”的快乐给予特殊待遇，这种做法看似不讲道德，甚至有些精英主义：“亲爱的傻子，如果你能够欣赏心灵的快乐，你也会舍弃啤酒，选择快乐的。”幸运的是，有一种观点能将这两方面统一起来，密尔曾毫不犹豫地将其抛弃，但我认为这个观点优于密尔的主要观点，也得到了较新的心理学理论支持。

芭芭拉·弗里德里克森在积极情绪的“扩大-建构”理论中提出，使我们感到愉悦的事情通常也是积累资源的方式。美味的食物能提供营养资源；与朋友在一起能积累社会资源；学习是在积累认知资源。在可持续、可共享的资源积累过程中，衍生出来愉悦似乎与密尔所说的“高级快乐”含义相符。这恰恰为密尔的“高级快乐”理论提供了一条更加有力的功利主义论据。

密尔想要表达的意思是，尽管更多人喜欢啤酒，但哲学依然优于啤酒。对此，他坚持认为既了解啤酒又了解哲学的人会选择哲学，因为哲学给人以更好的愉悦感，也就是“更高级的”快乐。在他看来，愚蠢的醉汉错过了上好的机会。但我想，密尔为精神生活和广义的高级快乐辩护的方式也许并非最佳。我认为，对傻瓜而言，做快乐的傻瓜（也许）更好；对于其他人而言，做苏格拉底更好。（是否有些耳熟？）同样，做快乐的傻瓜也许对现在的傻瓜更好，但对于未来，则未必有那么好。这样看来，我们并不能说阅读柏拉图会比喝啤酒带来的快乐更多。事实上，阅读柏拉图之所以更加优越，是因为阅读的快乐会带来更多其他的快乐，不仅是对阅读者本身，对其他人也是如此。密尔试图通过即刻获得的个人利益，（“来吧，这是更加愉悦的快感！”）为高尚的生活辩护。但他其实应该使用“更大范围的利益”为自己辩护：高级快乐之所以高级，是因为它们所造成的影响更加深远，与它们给人的感觉无关。

在密尔和边沁两种观点的统一过程中，仍存有一个问题：假定不同生活方式产生的长远影响相同，有人会认为充斥着性、毒品和摇滚的生活可能优于平静的冥想生活。对于这个观点，我的感触十分复杂。看到他人纵情酒色时，只要他们没有使自己或他人的境遇每况愈下，我很乐意说，“你喜欢就好！”但倘若让我设想自己是否能以这种方式生活，答案就没这么简单了。如果知道不论是自己还是他人，都无法再获得更大的发展，我是否会放弃舒适的教授身份，加入永不停止的派对呢？也许不会。但倘若是出于忘却烦恼、潜心做事的目的，我也许会去参加派对。

不管怎样，此处的重点是，我们不必把功利主义视为“卑劣的道德”。功利主义之所以将某些“高级”快乐置于“低级”快乐之上，是有其合理解释的。

高级快乐之所以更加高级（至少在某些时候），并非因为它们带来的快感更加愉悦，而是因为从长远来看，它们是更加有用的快乐。

## 我们关注谁的幸福？

我关注所有人的幸福。除了重视体验之外，功利主义的第二个定义性特征便是公平不倚，所有人的快乐同等重要。但是在功利主义的世界中，公平不倚并不意味着每个人都一定要同等快乐。正如北方牧民所说，如果一个世界中，不论贡献大小，每个人都获得同样的收获，那么这个世界必然死气沉沉，人们没有做事的动力。因此，将幸福感最大化的方式并不是规定每个人必须同等快乐，而是鼓励人们去做能将幸福感最大化的事。衡量道德的成功时，我们对所有人的快乐一视同仁；但成功的状态中也势必会包含物质财富和快乐分配不平衡的情况，这种不平衡不是我们理想的状态，但它依然是合理的，因为如果不是这样，整体情况便会更加糟糕。

“谁的幸福”这个问题还可以从另一个角度理解。我们可能会问：谁定义的幸福才是真正的幸福？对我来说，幸福就是两种冰激凌；对你来说，幸福就是阅读柏拉图；对别人来说，幸福就是被一位打扮成牧羊女的300磅的女人绑起来用鞭子抽。那么，谁的幸福呢？

事实上，这个问题更像是语言表达层面的问题。我们当然可以说不同人眼中的幸福各不相同，但这种说法只是在制造不必要的混乱。更清晰的表达应该是：幸福对所有人都相同，只是不同的人会因不同的事而感到或悲或喜，比如，两种冰激凌会让我感到幸福，但你不会……

我之所以认为这是语言表达问题，是因为有人会质疑幸福对所有人来说是否真的相同。怀疑全世界的人们对幸福的体验是否相同着实十分大胆。下面的句子描写的是18世纪一位居住在日本的小男孩儿：“桓武下到井边，惊奇地发现水又出现了，他感到非常高兴。”读完句子，你感到困惑吗？当然不会，因为你对这句话的理解完全到位，桓武的感觉与你自己开心时的感觉大致相同。再来看看这句话：“桓武发现两只死瓢虫被压扁在石头上，他感到非常高兴。”这太奇怪了。之所以会有奇怪的感觉，是因为你把自己对幸福的理解，完全应用到了桓武身上：显然，你从两种冰激凌中获得的体验与他从两只死瓢虫身上获得的体验是一样的，你和桓武之间存在文化鸿沟，你们两人的经历也可能大相径庭。但不管怎样，你们的体验依然存在共通之处，从某种程度上讲，它们都是积极的，或者都是消极的。因此，幸福才是我们所要找的通用货币。

## 如何衡量幸福

我们已经一致同意，世界各地的人们都有能力获得积极（或消极）的体验。下面我们来讨论如何衡量幸福。在过去的几十年中，衡量幸福作为一项复杂的任务，吸引了无数社科精英展开思考与探索。但我在这里提出的观点与高深的科学分析无关，我想说的是，衡量幸福并不困难，难的是按照我们心中的标准对幸福进行精确衡量。我们无法精确无误地衡量幸福，因此在实际生活中产生了很多巨大的困难，但这绝对不是什么深奥的哲学问题。

看看里卡多，他因膝盖骨受伤而住进医院，痛苦万分；看看比特丽姿，她斜倚在微微晃动的吊床（315美元）上，用iPad（499美元）阅读新闻。我们会认为，比特丽姿这一刻的感觉优于里卡多。但我们是如何判断的呢？可以通过询问：如果用数字1～10来描述你此刻的感受，你现在感觉如何？里卡多选了2，比特丽姿选了8。通过这种方式，我们只能测量出两人的相对幸福感。

这种测量方式合理吗？我们不知道。也许里卡多感觉不错，只是不想吓到我们而已；也许他正在经历人生中最难熬的阶段，但由于担心别人认为他抱怨太多，里卡多选择了2而不是1。也许比特丽姿正在经历内心的煎熬，但她不愿向我们承认，甚至自己也不愿承认；也许她只是不愿意给出过高的数字，所以对自己的幸福感做出了偏低的评价。我们询问里卡多和比特丽姿两人当时感觉如何，但我们也可以询问两人目前的总体生活状况如何。这样一来，幸福的衡量就变得更加困难：也许里卡多的生活其实一切顺利，比比特丽姿的生活还要好，只是在这一时刻，里卡多不这样认为。

这些都是严肃的问题，这些问题的存在并不意味着我们无法衡量幸福，只是意味着我们对幸福的衡量只能停留在估测层面。我们能否做出足够准确的估计取决于我们的衡量目的。如果我们想了解一个人究竟有多幸福，或者想要比较在相似状况下两人的幸福程度有何差异，这种估测也许不够准确。但幸运的是，作为社会整体，我们不必精确衡量社会中每个个体的幸福，再据此做出重大决定。相反，我们需要了解的只是整体趋势：何种政策能够提升幸福感？何种政策会降低幸福感？

这时，幸福新定义的优势便显现出来了。例如，我们知道，失业会对人们的精神状态造成沉重打击，由此引发的心理损失远远超过经济损失；但如果你已经十分富足，那么收入的略微减少并不会对你的幸福感造成太大影响。失业对有些人来说正是天赐良机；有些人也许只因年收入从22万美元降到20万美元就如坠地狱一般。但总体来说，经济数字的变化对幸福感的大体影响是可知的，我们也因此能够做出更加明智的政策决定，如在增加税收和创造就业之间进行选择等。尽管我们无法对每个个体的幸福感进行精确衡量，但这条结论是确切无疑的。

当人们为自己无法衡量幸福感而担忧时，他们心中也许还有其他考虑。我们并非没有想过直接询问人们的感受，但我们担心简单的询问也许不够。我们想要的是对幸福感“真正”的测量，能够排除主观影响的直接测量，就像温度计的测量一样，不受人们感知冷热的影响。随着脑功能成像技术的出现，这种测量技术的产生也许不再遥远。*但即使真的有人发明了神经幸福感测定仪，情况也难有大的改观。我们将仪器测量结果与人们汇报的结果进行比较，也能够发现出于各种原因而没有如实汇报幸福感的人。但在多数情况下，我们需要的只是对简单问题的一个简单回答。正如丹·吉尔伯特所说的那样，验光师不必用仪器扫描脑部，计算哪副镜片的视觉认知效果最为清晰；她只会直接询问：“现在看上去怎么样？”

衡量幸福并非难以逾越的难题，即使它能被称为难题，那也是摆在所有人面前的难题，并非功利主义者自身面对的难题。所有人都知道，我们做出的选择与幸福感是有关联的。因此，即使你不赞同功利主义观点，不认为幸福是最为重要的终极价值，只要你认为幸福在某种程度上是重要的，那么衡量幸福就是有必要的！

## 功利主义者总是“精于算计”的吗？

如果用一个词概括人们对功利主义的种种误解，那就是“精于算计”。人们心中“精于算计”的功利主义者形象有两个特点。

首先，“精于算计”的人是不好的、自私的，这种人永远在思考如何才能满足自己。将功利主义者看作“精于算计”的人实在冤枉，功利主义的理想是公平不倚。对一位理想的功利主义者来说，他人的福祉与自己的福祉同等重要——这是对黄金法则的最好注解。[注]功利主义绝非自私的哲学思想，它所面临的阻碍恰好相反：功利主义思想要求人们过于无私。（稍后对此还有更多讨论。）

然而，这种误解中有一点是正确的：即使是充满善意的道德计算，也可能引人误入歧途。起初，某人的谋划算计也许是为了更大范围的利益，但经过各种形式的自我麻痹后，这种谋划算计最终变成谋取私利的行为。（“一切为了爱罗马！”）事实上，谋划算计源于对道德机制的不信任，至少是暂时的不信任，也就是对第2章提到的优先考虑“我们”的社会本能持有怀疑态度。人们担心一旦放弃道德导航系统的指引，开始进行道德计算，便可能会惹祸上身。第2章末尾描述的公共财产游戏试验便印证了这一点：更多的思考使人倾向于不劳而获，放弃合作。

对道德计算的担忧与人们心中功利主义者的第二个形象特征相关：在人们心目中，功利主义者总在不停地进行道德计算。一个典型的形象就是：功

利主义者站在商店的走廊里，计算着从商店里偷窃的成本和收益。幸运的是，大多数人都不会进行这样的道德计算。这种行为似乎正是功利主义者所推崇的，但细想就会发现，这种行为恰恰与功利主义思维完全相反。为什么呢？因为频繁计算何种行为符合更大范围的利益这种行为本身就不符合更大范围的利益。倘若只要我们能说服自己，证明我们做的事符合更大范围的利益，我们便放心去做任何所想之事，那么世界将会陷入灾难。人类在为自己谋私利方面已然臭名昭著（参见第2章和第3章），但在考虑个体行为的长期、世界性影响方面却不很擅长。因此在日常生活中，听从道德本能的指引会让我们活得更好，最好不去思考小偷小摸这类行为是否符合更大范围的利益。人类的道德本能在生物学层面和文化层面都得以进化，帮助人类将“我们”的利益置于“我”的利益之上。但在日常生活中，我们却总试着通过计算摆脱本能，将自己置于危险境地。

从这一点出发，你可能会想，功利主义思想会不会早已备受攻击，成为过去：既然道德本能如此可靠，能够成就更大范围的利益，那么功利主义也好，其他道德哲学也好，还有什么用处？这里很重要的一点是，我们所提到的两个悲剧并不相同。再次强调，道德本能应对公地悲剧（“我”与“我们”之间的对立）十分有效，但对于常识道德悲剧（“我们”与“他们”之间的对立）却无能为力。功利主义的作用就是让道德本能引导我们绕过日常生活中的道德诱惑（“我”与“我们”之间的对立），在思考新草地上生活的问题时（“我们”与“他们”之间的对立），再采用清晰的功利主义思维。第五部分中，我会进一步介绍这种机制的工作原理。

本章开头，我们便将崇尚个人主义的北方牧民和崇尚集体主义的南方牧民放在一起，将两方的理想和言论进行比较。如你所想，功利主义者并不一定偏向集体主义，也不一定会偏向个人主义。我们有个绝妙的想法，就是请不同部落的牧民暂时搁置己方意识形态，认真考虑何种方式更加有效；何种方式能将新草地上的幸福感最大化；最有效的方式究竟更倾向于个人主义还是集体主义。找出最有效的方法需要我们搁置偏见，收集信息，评估不同的政策与行为在真实世界中的进展程度。如前所述，功利主义就是只留下基本原则的实用主义。**

## 如何将幸福感最大化，由谁来决定？

至此，我希望你已经掌握了功利主义思维的窍门儿，可以自己找出这类问题的答案。不过为了善始善终，我们再来回答最后一个问题。

从功利主义的角度出发，选择何人握有决定权不过是一个普通的选择，与其他选择并无异处。并不存在官方认定的功利主义决策者，他也不会拥有炫目的头衔。功利主义者认为，一个好的决策系统中，决策者是做出决定

产生最佳后果的最佳人选。从理论上讲，将决策权集于一身的哲学家国王也许符合这样的标准，但我们所了解的历史和人性都告诉我们，这是不可能的。相反，在代议民主的制度下，辅以自由的媒体、广泛的教育等，我们似乎生活得更好。

## 总结

功利主义将两个合理且得到普遍认可的观点相结合，这两个观点可以回答以下两个问题：真正重要的是什么？真正重要的是谁？

根据功利主义思想，个人体验的质量最为重要。功利主义并非偏爱世俗实用的事物，排斥细微之事，其目的不是为了在洗衣房层面上将“功用”最大化。反过来，功利主义也并非偏爱“我们最喜爱的事”等细微之事，排斥意义更加深远、更加重要的事情。功利主义对我们所看重的几乎所有价值都表示赞同，从家庭、朋友、爱等与人际关系相关的价值，到诚实、坚韧等个人美德；从真理、艺术、运动等高尚的追求，到自由、公正等治国之道。但功利主义者认为，一切价值观的价值，都源与它们对人类体验的影响。**这个想法也许正确，也许错误，我们尚未来得及考虑与其对立的观点，但这个观点看上去似乎合情合理。更重要的是，每个有思想的人在暂时搁置部族思想时，都能够理解这个观点并能对此表示赞同。

功利主义思想的第二个要素是公平不倚，这是道德的普适本质，在黄金法则中也得到了提炼。有了这个要素，我们便可以将功利主义思想概括为：一切以幸福为上；每个人的幸福同等重要。每个人并非一定要同等幸福，但确凿无疑的是，不同人的幸福从本质上讲并没有高下之分。

幸福是可以被衡量的，尽管精确衡量十分不易。通常情况下，也无须精确衡量个体幸福，我们只需要通过研究人群的幸福，总结出能够提升或降低幸福感的行为，便可以得到我们想要的关于幸福的信息。

显而易见，我们无法找到能在长期内将幸福最大化的因素，有人认为这是功利主义的致命缺陷。但仔细想来，这种观点完全站不住脚。在寻找长期影响最好的事件过程中，所有人（除了对长期影响漠不关心的人）都需要猜测，也许是有根据的猜测，或是其他形式的猜测。功利主义的独特之处并不是对长期影响的考虑，而在于它将长期影响作为最终的决定因素。

从最基本的层面来讲，功利主义并非决策过程，而是一种理论，揭示了基本层面上最为重要的因素，指出了什么值得我们重视以及为何值得我们重视。功利主义并不要求我们对行为的成本和收益进行频繁计算，相反，它要求我们在多数情况下相信道德直觉，因为道德直觉通常比频繁的道德计算对我们更加有利。

功利主义不要求我们追随领袖，即使领袖声称自己代表更大范围的利益。相反，功利主义要求我们以可能会导致好结果的方式做出决策，充分考虑人性的局限与偏见。历史上的乌托邦政治提醒我们，面对声称自己能实现更大范围利益的领袖，功利主义要求我们保持怀疑之心。

总之，功利主义将黄金法则的公平不倚与人类体验的通用货币相结合，建立了允许权衡道德的道德系统，并提供了权衡裁决的方法，所有部落的成员都能够理解这种行事方式。

## 重要的趋同现象

从一万英尺的高空俯视新草地，看着道德机制和道德直觉各不相同的部落争斗不休，功利主义的务实解决方案似乎如此显而易见：人们应当将各自部落的意识形态暂且搁置，共同找出新草地上最好的生活方式，然后按照这种方式生活。针对第一部分提出的道德问题，这就是我们分析得出的结论。但正如前文所述，第二部分的心理分析中提出了另一种推理思路。

前文提到，人类的大脑具有双加工机制，既有提高思维效率的自动模式，也有使思维更加灵活的手动模式。将人类大脑比作可在两种模式间切换的相机是有意义的，因为这个比喻不仅准确描述了人类道德心理，而且为现实中的难题提供了答案。当代牧民如何解决争议？前文中，我们将这个问题转化为哲学问题：何种哲学能够作为元道德？但我们也可以从心理学角度看待这个问题：何种思维最适合新草地上的生活？对于这个问题，照相机的比喻十分有用。

摄影时用哪种模式好？自动模式还是手动模式？毫无疑问，两种模式都不能在绝对意义上达到最好。对于不同的摄影对象，两种模式各有所长。典型的摄影场景下，即相机制造商提前设计好的场景中（“人像”、“景物”），自动模式基本就能满足你的需求——只需瞄准，然后按下快门就好。但如果你面对的场景相机制造商从未想过，或者你的审美偏好与相机制造商不同，那么你需要的大概便是手动模式。

于是问题出现了：从道德的角度看，我们现在身处哪种场景？新草地的问题需要使用自动模式还是手动模式进行解决？

公地悲剧的避免是通过一整套自动设置实现的——由道德情感在一定的群体内促进并维持合作。但常识道德悲剧本身便源于自动设置：因为不同部落的自动设置各有不同，他们看到的世界经过了不同道德镜片的过滤。公地悲剧是自私的悲剧，但常识道德悲剧却是道德僵化的悲剧。新草地上之所以会有冲突，并不是因为牧民们无可救药的自私、邪恶，甚至善恶不分；而是因为他们无法走出各自的道德视角。那么牧民们应当如何思考？

答案已经浮出水面：他们应当切换到手动模式。

这意味着什么？第4章已经提供了一些线索：手动思维模式和功利主义思维模式之间，似乎存在着某种联系。**处理人行天桥困境及其他类似情况时，手动模式建议我们尽可能多地拯救生命，直觉却让我们做出相反的行动。支持功利主义思维的主要大脑区域是背外侧前额叶皮层，它同时也负责指挥我们在其他领域灵活处事，比如坚持节食、降低种族主义倾向等。面对道德困境时，反对功利主义思维的主要大脑区域是杏仁核和腹内侧前额叶皮层，**面对族群以外面孔等陌生事物，永远保持高度警惕的也是这一部分区域。这并不能证明功利主义思维是正确的，也不能说明非功利主义思维是错误的。稍后我们会发现，大脑的手动模式同样能够应用非功利主义原则，我们也不愿用“神经联系的错误”盲目指责道德直觉。但不管怎样，这确实是一种值得注意的趋同现象。

如果我是对的，那么从某个视角看似正确的道德哲学和从某个视角看似正确的道德心理之间这种趋同一致的关系就并非偶然。如果我是对的，不论是从哲学层面还是心理学层面来说，边沁和密尔所做的事情与之前的研究便有了本质的差别。他们超越了常识道德的局限，将道德问题（几乎）完全切换到了手动模式。他们将刻板的自动模式放在一边，提出了两个非常抽象的问题：第一，真正重要的是什么？第二，道德的本质是什么？他们总结出，人的体验是真正重要的，而道德的本质则是公平不倚。将这两个观点融入手动模式，我们便得到了功利主义观点：应当最大限度提高个人体验的质量，将每个个体的体验看得同等重要。因此，早期的功利主义者采用以含义抽象而著称的黄金法则**，体现了公平不倚的思想，然后添加通用道德货币，即个人体验的概念，使其观点更加具体。

但个人体验真的是恰当的通用货币吗？这真的是最好的哲学观点吗？如上所述，功利主义面临很大的争议，很多专家学者都认为功利主义思想问题重重。上文提到，功利主义在某些情况下似乎会给出错误的答案：即使将人推下天桥能够产生更好的后果、增加整体幸福感，这种做法似乎依然不对。反对功利主义的诸多声音中，这只是直觉上让人震撼的例子之一。我们会在第四部分对这些异议详加论述。但在此之前，让我们先对通用货币这个概念展开进一步讨论。是否还有其他哲学思想能够弥合“我们”和“他们”之间的裂缝？这些哲学思想是否优于功利主义？是否存在绝对正确的哲学思想，也就是说，是否存在道德真理？如果答案是肯定的，那么功利主义是这个绝对正确的思想吗，抑或是其他的思想？接下来的两章将会讨论我们所面临的选择（第7章），阐释为何只有功利主义思想才能够作为当今世界的元道德（第8章）。

---

1. 你们愿意人怎样待你们，你们也要怎样待人。《马太福音》7：12；

《路加福音》6：31。——译者注

# 第7章 寻找通用货币

民主要求笃信宗教的人们将其关切从具体的宗教价值移开，转而投向普适价值。民主要求人们有权对提议进行讨论，并据理进行修改。出于宗教信仰，我也许反对堕胎，但如果我想通过立法禁止堕胎，就不能简单地搬出教堂的布道或引用上帝的意愿。我必须说明为何堕胎违背了所有人的原则，包括拥有各种信仰的人，也包括完全没有信仰的人。

——巴拉克·奥巴马

正如奥巴马所说，现代牧民需要通用货币，即用于衡量不同部落间价值的通用标准。没有通用货币，便没有元道德，进行道德妥协和得失权衡便没有依据。寻找通用货币是艰巨的任务，有些人甚至认为这是天方夜谭。

最根本的阻力来自部落的效忠派。奥巴马敦促有信仰的道德思考者将其关切从具体的宗教价值移开，转而投向普适价值。但倘若有人坚定不移地相信自己的信仰才是普适道德真理的源头，那该怎么办？这种情况下，区分普适价值和具体宗教价值毫无意义。（奥巴马也意识到了这个问题。*）里克·桑托勒姆（rick santorum）是一名保守派议员，在2012年总统选举中成为共和党候选人。他曾声称，奥巴马的立场让他感到恶心。“我们身处的是怎样的国度？竟然只有无信仰的人才能够站在公共广场发表看法。”*桑托勒姆的话有些夸张，并没有人规定有信仰的人们不能发表看法。相反，奥巴马提出，有信仰的人应当以世俗的方式提出自己的道德观点。但对很多宗教道德家来说，这项要求就像是让一位芭蕾舞蹈家身穿带垫子的相扑运动服跳舞。试问“同性恋是憎恨上帝的生活方式”这句话如何用世俗的方式表达？难怪桑托勒姆会对此感到恶心。

前文提到，寻找通用货币的另一个阻力来源于众所周知的“相对主义者”、“共产主义者”以及其他不相信普适价值的人。他们认为，通用的道德货币根本不存在，那些认为通用道德货币存在的人与宗教基要主义者一样，都在向世界推销自己的价值观。就像汽车保险杠上的车贴写的那样：“一切道德都是本土化的。”

此外，当代道德家也提出了质疑，他们对通用的世俗道德十分看好，却对我所提倡的这种道德不甚乐观。前文提到，很多现代道德思想家认为道德的本质与权利有关。他们认为，世俗普适的道德真理从本质上讲是关于谁拥有何种权利、权利的优先等级如何的问题。持有这种观点的不只是哲学家，当别人要求我们证明自己的道德信念，或让我们提出能够“接受讨论，据理修改”的方案时，我们往往也会诉诸于此。比如，我们为堕胎辩

护时会提到女性“选择的权利”，或是胎儿“生存的权利”。我们坚持认为某些权利比其他权利更加重要，有时甚至会否认其他权利的合理性。

同样，功利主义者也会谈及权利，也会将不同种类的权利进行比较：如果保留选择权，牺牲生存权能将幸福感最大化，那么“选择权”便是优于“生存权”的。但这与大多数人对权利的思考方式并不相同。请回想小火车的问题：将人推下天桥也许会将幸福最大化，但这种做法似乎依然是对个人权利的无情践踏。按照通常的理解，权利并不能被“换算”为后果，它们永远“胜过”后果。

如果我们（应当）拥有的权利事实与产生最好结果的行为事实不相符，那么前者属于怎样的事实呢？道德事实的传统模型之一是数学：第101个质数是多少？我们不知道。但可恶的是，如果我们想知道，就可以通过计算得到。同样，如果我们努力思考道德，也许就能根据基本原理想出道德事实，那样我们就会得到完全不同的一种通用货币：对权利的存在以及不同权利之间轻重缓急关系的事实描述。我们就可以据此判断选择的权利是否优于生存的权利，就像我们计算第101个质数一样。当然，没人认为道德事实等同于数学事实，可以通过数学计算得出结果。这个类比的意义在于说明，道德事实与数学事实都是抽象的真理，但通过足够深入、客观、认真的思考，我们便可以找到真理。对很多现代道德思想家来说，这就是他们的梦想。

道德事实的另一个模型来源于自然科学：有些部落认为地震的起因是一条巨大的鲶鱼扭动身子蹦跳；有些部落从则认为地震是大地生病颤抖的结果。可是科学告诉我们，地震是地壳的大型板块漂浮在熔岩之上，相互摩擦的结果。现代科学对于地震成因的解释并不是另外一个部落神话，而是建立在证据上的解释。如果有足够的时间和耐心，所有部落的成员都能认可这种证据。从广义上来说，科学为我们对自然界的理解提供了一种通用货币。比如现代板块构造理论已经得到科学家的广泛认可，不论他们生活在哪个大洲，文化背景如何。考虑到这一点，有人便希望科学能够揭示道德背后的本质，给道德下定义，而不仅仅是进行描述（就像第2、3、4章那样）。也许科学能告诉我们哪些权利是真实存在的，它们之间的轻重关系如何，排出一份“道德元素周期表”。这种方式也可以为我们提供所需的通用货币。

本章将讨论寻找道德通用货币的途径，以及各种途径所对应的元道德。考虑到具体的寻找过程，有两种方式可供选择。如果我们想在纯粹的理论层面有所建树，便可以对道德真理进行探寻，也就是寻找普适原则，探究新草地上的牧民应当如何生存，我们究竟拥有哪些权利，负有哪些责任。

由此，我们将依照上述三种探寻道德真理的方式开始讨论，即宗教模型、

数学模型以及科学模型。我将阐明为何三种模式都不可能满足我们的要求。如果我们无法从上帝、推理以及自然等外界因素中找到道德真理，那么就只能转向更加朴素的元道德，寻找适当的部落间机制。不论它是不是道德真理，我们都只能安于此道。**下一章将采用第二部分提到的道德心理双加工理论，论证为何功利主义是唯一能胜任这项任务的思想。

## 通用货币来自于上帝吗？

对很多人来说，普适道德规则只有一个来源，那就是上帝。然而，要想寄希望于上帝的道德权威，我们首先面临至少两个主要问题。一是上帝权威的范围问题，二是对上帝意志的感知问题。

第一个问题要追溯到柏拉图，他对道德权威与神的意志之间的关系提出了质疑。将柏拉图的问题用现代神学语言翻译出来就是：坏事是由于不被上帝认可才成为坏事吗？抑或是上帝之所以不认可坏事，是因为这些事本身是坏的？以强奸为例：强奸是坏的，上帝和我们都这样认为。但上帝最初是否有可能认为强奸是好的？上帝是否有能力把强奸变成道德上可以被人接受的事？如果你赞成柏拉图的想法，认为答案是否定的，那么你的观点其实是：上帝没有制定道德规则，至少没有制定全部的道德规则。有些道德规则，包括禁止强奸等十分重要的规则，是不受上帝意志控制的。倘若如此，那么我们便需要通过某些世俗的方式来解释为何一些事情是正确的，另外一些则是错误的。或者，你会认为上帝确实能够按照自己的意愿制定道德规则，比如撤销对强奸行为的禁止，甚至鼓励强奸。但如果上帝的行为如此随心所欲，是非界限如此模糊，那么凭什么说上帝的意志就是道德的呢？如果我们只能无条件接受上帝的意志，只能以“因为他是这样说的”作为解释，那么上帝的意志不过是一系列随意制定的规则，不过是碰巧得到了至高权力的支持而已。

柏拉图的观点已经存在了很长时间，但他的观点并不能将宗教道德完全摒弃。一方面，柏拉图的观点在今天已经不像当初那样令人信服。柏拉图身处多神崇拜社会，古希腊诸神本身就是无秩序的混乱群体，很难将他们视作道德楷模。因此，道德独立于诸神意志这个观点便很容易被人接受。但神学领域有一个更加现代、更加复杂的概念，可以反驳柏拉图的观点。现代神学家可能会说，道德与神的意志不可分割，上帝确实无法赞同强奸行为，但这并非因为上帝的力量有限。上帝的力量无边无尽，甚至超越了空间和时间，他的行为不是可以用成功或失败衡量的谨言慎行，而是一种现实特征的代表，人类智慧有限，我们的理解只是管窥蠡测。上帝无法赞同强奸行为恰恰印证了其意志本质的永恒和完美。（对无神论者来说还不错，是吧。）简而言之，心思缜密的神学家可以用质朴的上帝概念，将柏拉图的观点驳倒（至少他们自己这样认为）。在我看来，神学家的观点像

是在说“这是神迹”，只不过表达方式让人感到眼花缭乱而已。我们不必马上解决这个问题，对我们来说，神的道德真理面临一个更加严重的问题：在何为道德真理这个问题上，我们无法得到直接的、触及问题本质的答案。

假设我们认可上帝存在，也承认道德真理来自他的权威意志，那么我们如何感知上帝的意志呢？比如，很多基督徒认为同性恋不道德，如何证明呢？第一个证据出现在《圣经·旧约》中，《利未记》18章22节写道：“不可与男人苟合，像与女人一样，这本是可憎恶的。”接下来是《利未记》20章13节：“人若与男人苟合，像与女人一样，他们二人行了可憎的事，总要把他们治死，罪要归到他们身上。”当然，这些句子还需要进一步解释，但按照很多人的理解，我们姑且认为这些句子明确表达了对同性恋的谴责。但刨根问底的道德家们需要考虑，我们是否应当认真对待这些谴责。因为《圣经·旧约》中谴责的很多事在今天看来已是司空见惯，而其中所宽恕的很多事在今天看来却是十恶不赦。

下面是一封写给劳拉·施莱辛格博士的公开信，反映了《圣经》理解中存在的问题。劳拉是一名保守派评论员兼广播主持人，她曾经引用《圣经·旧约》作为依据，批判同性恋。

> 亲爱的劳拉博士：
>
> 感谢您不遗余力的努力，教导人们学习上帝的法律。我从您的节目中学到了很多，并且还在向尽可能多的人传播这份知识。比如，有人试图为同性恋的生活方式辩护时，我只需提醒他，《利未记》18章22节中清晰表明，同性恋是令人厌恶的。无须多言。
>
> 然而，涉及某些具体的规则，以及如何遵从这些规则，我迫切需要您的建议。
>
> 在祭坛上焚烧公牛作为献祭时，我知道上帝很享受这种味道（《利未记》1：9）。但问题在于我的邻居，他们声称这种味道并不好闻，我是否应当责备他们？
>
> 我想将我的女儿卖作奴婢，就像《出埃及记》21章7节中所赞许的那样。在当今时代，您认为她的标价定位多少比较合适呢？
>
> 我知道在女性每月一次不洁净的周期中，我是不能接触她的（《利未记》15：19-24）。问题在于，我怎么能知道她们何时处于周期当中

呢？我曾试着问过，但大多数女性都会被这个问题惹恼。

《利未记》25章44节提到，我可以拥有男性和女性的奴隶，只要他们是从外国买来的就好。一位朋友说这条规定适用于墨西哥人，但不适用于加拿大人。您可以为我解答这个疑团吗？我为什么不能拥有一位加拿大奴隶？

我的邻居坚持要在安息日工作，《出埃及记》35章2节清楚地写明，他应当被处死。从道义上讲，我有义务亲自处死他吗？

一位朋友认为，虽然食用海鲜贝类也是可憎的行为（《利未记》11：10），但总比同性恋要好一些。我不认同他的观点。这个问题您怎么看？

《利未记》21章20节说，如果眼睛有毛病，便不能靠近神坛。不得不承认，我戴了一副近视镜，要想靠近神坛，我的视力必须达到20/20吗？或者这中间还存在着一定的回旋余地？

我的大部分男性朋友不顾《利未记》19章27节中的训诫，都剪短了头发，剃掉了鬓角。他们该怎样被处死呢？

我从《利未记》11章6节到8节中得知，触摸死猪皮会让我沾染不洁净。可如果我戴上手套，可以继续打橄榄球吗？

我叔叔有一个农场，他违背了《利未记》19章19节的训导，因为他在同一块地上种了两种不同的作物。他的妻子也没能遵守规定。因为她穿的衣服是用两种不同的衣料制成的（棉和涤纶混纺的）。他还经常口出恶言，说亵渎的话。我们真的有必要把全镇的人召集起来，用石头将他们打死吗？（《利未记》24：10-16）我们能不能简单地在私下家族会议上把他们烧死，就像处死与姻亲通奸的人一样？（《利未记》20：14）

我知道您对这些问题已经有了深入的研究，相信您一定能够帮助我。再次感谢您提醒我们，上帝的言语是永恒不变的。

您忠实的信徒、仰慕您的追随者

J·肯特·阿什克拉夫特

如你所想，这场争论并没有就此罢休。思维细密的《圣经》解读者们总可以说，《圣经》的某些话语可以作为直接道德准则，另一些则不行。但阿

什克拉夫特先生的一个观点无疑是正确的：即使是在单一宗教传统的背景下，要想建立道德真理，简单引用《圣经》原文也是远远不够的。不同宗教传统为《圣经》的权威解读问题争执不下时，问题就更加复杂了。如果我们想要通过经书原文建立道德真理，我们就必须决定，具有真正权威的是哪个宗教传统的哪个文本，我们需要对哪个段落进行何种解读。以宗教为载体的道德争议中，处于对立面的双方几乎不可能在权威的文本类型、段落选择以及阐释方式等问题上达成一致，因此，以经书文本为依据的观点基本只能用于解决最狭义的学术道德争议，除此之外，就无能为力了。

同样的问题也存在于人们从梦境、幻觉、神迹等其他形式的人神交流中所获得的道德真理上。奥巴马在同一次演讲中说道：

上帝命令亚伯拉罕献出自己唯一的儿子，他没有争辩，将以撒带到了山顶，骗到祭台前，把刀举起，准备遵从上帝的命令……但我想，所有人离开教堂后，如果在屋顶上看到亚伯拉罕举着刀子，至少我们都会给警察打电话，然后希望儿童及家庭服务部门能将以撒从亚伯拉罕身边带走。我们之所以会这样做，是因为即使亚伯拉罕的经历真实可信，我们却没有听过亚伯拉罕所听到的声音，也没有见过亚伯拉罕所看到的情景。因此最好的选择就是按照我们所听到看到的事实行事，不管这种规则是普通法律还是基本的理性。

不管怎样，也许任何证据都不能阻止效忠派听从部落召唤；也无法说服桑托勒姆议员和劳拉博士，让其相信如果宗教信仰不能用世俗的语言表达出来，便无法作为公共政策的依据。我们至多能做的，不过是提倡中庸之道，提醒部落的效忠派，让他们意识到自己的行为并非以“常识”作为基础，相反，他们做的只是将各自部落内部的道德真理强加于人，但别人并没有听过他们所听到的声音，也没有见过他们所看到的景象。

我并不想针对宗教道德真理是否存在发起辩论，我想表达的只是，也许道德真理取决于不同部落对上帝意志的不同解释，但我们的任务是寻找通用货币，在这件事上，上帝帮不了我们。（有鉴于此，便不难理解为何自省行为不鼓励人们信仰上帝。）

世上各种信仰有很多相似之处，它们都教导我们与邻为善、不要撒谎、不要偷盗、不要为自己的道德错误寻找借口。也就是说，世界上的各种宗教都能使其信徒幸免于公地悲剧，将“我们”的利益置于“我”的利益之上。宗教——至少是绝大多数宗教——所力不能及的，是让我们避免常识道德悲剧。宗教非但不能缓解“我们”与“他们”之间价值观的冲突，反而会造成冲突加剧。要想寻找通用货币，我们需要另觅他法。

## 道德与数学相似吗？

关于信仰的讨论到此为止。下一个话题是：推理。我是推理的狂热推崇者，这本书，乃至我毕生事业的目的，都是为了寻找更加理性的方式来理解道德。但我认为，关于道德的理性主义观点中有一种是过于偏激的，这类态度强硬的理性主义者认为，道德就像数学一样：道德真理是抽象的真理，只要思维清晰，便能找出这个真理，就像数学家们探寻数学真理一样。比如，康德曾提出一个著名观点：诸如不能撒谎和不能偷盗等真正的道德真理，可以通过“纯粹的实际推理”原则演绎而成。今天已经很少有人对这一观点表示明确支持。但不管怎样，当我们坚持认为自己的道德观点基于推理，与对方不同时，我们的思维中大多存有康德式强硬理性主义思想的痕迹。很多人或者声明，或者暗示，都会表明其道德对立方的观点存在推理缺陷，**就像2＋2＝5的道德等式一样。

如果道德与数学类似，需要满足什么条件呢？如何才能将道德完全建立在推理之上？数学家的职责是证明定理，所有的证明都始于假设。数学中的假设主要来自两个渠道：已被证明的定理和公理。公理是大家公认的、不证自明的数学表达。例如，欧几里得平面几何中有一条公理，任何两点之间都能以直线连接。欧几里得并没有对该陈述进行论证，他只是认定该陈述是正确的，并且认定你也能看出这句话是正确的。因为所有定理都是由已经证明的定理和公理推导而来，又因为定理无法无限回推，所以说所有数学真理最终都来自公理，来自于不证自明的基本数学真理。

如果道德与数学相类似，那么在论证过程中引用的道德真理就必须来自道德公理，来自一系列不证自明的道德真理。*用数学模型描述道德的最根本问题在于，经过几个世纪的尝试，没人能够找到一套适当的道德公理。适当的道德公理需要：（1）不证自明；（2）可以推导出大量道德结论，用于解决现实世界中的道德纠纷。***现在你可能会认为，我们显然无法用公理表述道德，因此道德与数学并不相似。但这个显而易见的结论会带来怎样的影响？这是一个值得我们停下深思的问题。

我们用堕胎的问题作为例子，第11章还会对此深入讨论。胎儿的生存权利比女性的选择权利更加重要吗？宗教教义无法解决这个问题。（应当采用哪条教义？以何种方式阐释教义？）我们也暂时不愿从纯粹功利主义的角度考虑权利问题。支持生命权的人认为，胎儿与天桥上的人一样，都拥有绝对的生命权利。（本观点不以堕胎的净成本和净收益作为评判标准。）支持选择权的人则认为，女性拥有绝对的选择权。这样的辩论将如何收场？

毫无疑问，推理无法解决这场争论。“推理”在这里指的是单纯推理的力量，或者像康德所说的“纯粹的实际推理”。前文提到，我是推理的狂热追随者，也渴望用推理解决道德问题。但如果道德与数学并不相似，那么只依靠推理便是不够的，通过推理并不能得出我们应当拥有何种权利，也无

法为权利的重要性排序。当然，这是因为一切推理都需要前提。如果乔的前提假设并非不证自明，而简刚好不喜欢由这个前提推出的结论，那么简便可以直接对乔的前提假设进行驳斥，同时驳倒其结论。

我想表达的意思是：如果没有不证自明的前提，纯粹的推理并不能回答我们的问题。推理只能迫使我们的真实信仰和道德信仰更加一致，这一点是很重要的。（第11章会继续讨论这一点。）数学推理能告诉我们如何对待$439569>3\times17\times13$这个算式，但道德推理却无法告诉我们如何看待堕胎这个问题。这是由于数学推理始于一小部分公认的、不证自明的假设，而道德推理则始于大量互相关联的、未经查证的假设，尽管假设的提出者认为它们非常合理，但只有很少一部分假设是真正不证自明的。（换句话说，道德认知论是融会贯通而非基础发展式的。）

作为一个聪明人，你可能希望借助推理将相互对立的人类价值观这团乱麻梳理清楚。遗憾的是，事实并非如此。但人们依然会将推理作为幌子，说道："与你的观点不同，我的观点是基于推理得出的。"事实上，这样的表述至多只能算是一半正确。如前所述，推理不仅能使一个人内心的道德观点更加一致，也能使部落内的观点更加一致。同时，下一章将会提到，将功利主义作为元道德的论证过程中，人类共有的推理能力扮演了十分重要的角色。但推理本身无法告诉我们应当怎样在不同部落相互冲突的价值观之间进行取舍；也不能揭示我们所拥有的权利，无法比较相互冲突的权利孰轻孰重。要想寻找通用货币，我们还需要另觅他法。

## 科学能揭示道德真理吗？

如果宗教和纯粹的推理都不能解决道德争端，也许是时候将目光投向我们最喜爱的、客观公正的事实来源：科学。或许科学能帮助我们提出道德假设。尽管这些假设并非不证自明，但却显而易见，人类现有的关于自身和整个世界的知识也能对此提供支持。

第1章概述了道德的广义科学理论，这个理论反映了达尔文以后人们逐渐形成的共识：

道德是人类不断对心理做出适应性调整的产物，它能够使性本自私的个体享受到合作的果实。

假设这个理论是正确的，（如前所述，这已经是我们仅有的选择。）既然我们已经对道德的自然功能有了一定的理解，那么面对道德真理的问题，我们是否拥有优势？如果道德的功能是促进合作，那么为何不能将最能促进合作的元道德认作道德真理？这个想法十分诱人，但我们仍需面对一些重要问题。

如果我们是对的，道德的进化的确是为了促进合作，但这个故事还有另外一面。从生物层面上讲，道德进化是为了促进种族内部的合作，其目的是在种族之间的竞争中占得上风。*自然选择之所以会偏爱促进合作的基因，唯一的原因便是：善于合作的个体更有能力在竞争中胜出。这种现象与其他所有的生物学调整一样，强调了道德的另一个更加普遍的功能，也是其最本质的功能：传播基因。进化的目的本身不是促进合作，合作之所以得以促进，完全是因为这种行为能将合作者的基因繁衍下去。进化也许偏爱与邻为善的人，但同时也可能偏爱有种族灭绝倾向的人，其背后的原因完全一致。因此，如果我们希望通过进化找到道德真理，那么我们确实是找错人了。（请注意，尽管上述论证过程仍存有争议，但其同样适用于文化进化。**）

从进化角度寻找道德真理时所遇到的问题引出了另一个更加普遍的问题，即“应然与实然的问题”，有时也被称为（或许不太恰当）“自然主义谬误”。*这个谬误依靠“什么是对的或者好的”来判断“什么是自然的”，主要由所谓的社会达尔文主义者提出，他们将自然界冷酷无情的竞争和优胜劣汰的法则作为人类社会的模型。今天我们已经知道，自然选择既可以促进友善的行为，也可以促进恶毒的行为（参见第2章），所以从进化论中寻找道德真理的灵感并不是法西斯行径。但我们不能因为某种行为符合道德的进化功能就认定其是正确的，也不能因为某种行为不符合进化功能就说它是错误的。同样，如果某件事符合进化的目的，也不能说明它是正确的。

事实上，从进化功能中寻找道德真理与“道德与数学相似”的观点是一致的，不过是换了一种表达形式。这种方式的背后隐藏着一条道德公理：最能满足道德进化目的的便是最好的。这条公理并非不证自明，也没有任何科学证据的支持，不过是一条假设而已。要想证明这条假设的模棱两可，试着用它解决争执就可以了。假设我们发现，以生命权为重的堕胎政策能帮助我们传播基因，如果是选择权的拥护者，这项发现会改变你的观点吗？当然不会，也不应该。你所相信的是女性的自由选择权，而不是传播人类基因。同样，如果进化事实得出相反的结论，以生命权为重的人也不应放弃自己的观点。（从这个角度来说，拥护选择权和拥护生命权的人可以一同到诊所外面呐喊：“进化，进化！传播基因不能代表一切！”）

你可能认为，从道德的自然功能角度寻找道德真理有一定的可行性，那么我们便对此进行探讨。我们将传播基因这个并不很道德的目标舍弃，而将合作这个更加切题的目标作为重点。合作是终极道德善行吗？在没有证据和进一步论证的情况下，我们能将其假定为终极道德善行吗？想象一下新版《星际迷航》中来自星际的威胁：博格人。首先普及一下，这里的博格并不是瑞典网球冠军，而是由神经机械个体（半机械，半有机）形成的族

群，他们是被同化的人类或人形外星人。博格人就像是高科技的超级智能蚁群，吞并其他生命形式，将其纳入自己的移动蚁巢。关于博格人有两点值得一提：一是他们是高度合作的；二是成为一名博格人似乎并不是什么好事情。可是，如果合作真的是终极道德善行，那么博格人的胜利似乎是宇宙生命的最好归宿，即使他们的生活十分悲惨，只要其社会高度合作，这种生活便是最好的选择。

博格人的例子提醒我们，将合作视为终极道德善行这一观点简直令人难以接受。但如果合作不是最重要的，最重要的又是什么呢？有人可能会说，合作的重要性并不是作为终极目标而体现的，而是体现在因合作而产生的好处之上。因为合作最终能够使人幸福、免于受苦，所以合作是至关重要的。这个观点听上去非常棒。

但这个观点距离最初的进化论已经太过遥远。经过一系列的调整，我们已经把道德真理的进化论变成了先于达尔文出现的道德论。也就是说，当我们终于将道德真理的进化论修改完毕时，此进化论已经不是彼进化论了。与其从进化的“价值观”出发，然后根据我们的需求不断修改，还不如直接从自己的价值观出发，从那里寻找通用货币。

## 备用计划：寻找共同的价值观

如果我们能够十分确定地识别上帝的意志，或者能由不证自明的基本规则演绎出大量的道德真理，抑或是我们可以像发现地震成因那样发现道德真理，那么我们便不必如此为难。现在我们不得不再次回头面对这个相互矛盾的道德价值泥潭。（后文简称“泥潭”。）

道德真理是不是根本不存在？我认为这个问题是不可知的。曾经一度，我认为道德真理的存在就是我们要考虑的问题，但我逐渐改变了看法。真正重要的问题是，我们是否拥有直接、可靠、确定无疑的方式寻找道德真理。我们需要的是在泥潭中清出一条小径，而道德真理是否存在其实并没那么重要。鉴于以上的分析，我非常确信，我们无法寻找道德真理。（如果确实存在一种解决道德冲突的权威方式，不依靠上帝的启示，不依靠纯粹的推理，也不依靠实证调查，那就说明我太过孤陋寡闻。）一旦我们放弃了泥潭中的挣扎，道德真理这个问题也就失去了实际意义。

（概括说来，关于道德真理的问题其实是，道德信仰面对现实做出了尽可能多的妥协和发展后，我们该对其进行怎样的描述？存留下的部分是否能叫作“道德真理”？或者只能算作是“剩下的部分”？我不再认为这一问题能够得到明确的回答，也不认为这个问题非答不可。***）

从泥潭中抽身出来，我们已经别无选择，只能利用共有的价值观，从这里

寻找通用货币。然而，确定共同的价值观并不像看上去那么简单，因为语言，特别是美好的语言，常常会误导我们。比如，两个家庭都很重视“家庭”这个概念，但这个概念也许并不能作为道德共识。如果要在工作中推广“家庭为重”的政策，这种含义下的“家庭”也许能够作为两个家庭的共同价值。但如果要解决“你的孩子对我的孩子做了什么”这个问题，两个家庭对“家庭观”的重视可能只会让事情更加糟糕。因此，“家庭”等抽象的道德观念会造成共同价值观方面的错觉，“自由”、“平等”、“生活”、“公正”、“公平”、“人权”等概念都存在同样的问题，后文还会进行解释。寻找真正的共同价值并不像乍看之下那样简单，因为深层的道德分歧常常隐藏在共同的道德名称之下。这样看来，真正的道德共识究竟在哪里？

也许你现在已经意识到，我认为功利主义价值观才是人类真正的道德共识。如前所述，我们都有过积极或消极的体验，都有过幸福或悲伤，也都意识到在道德的最高层面，**我们必须做到公平不倚，这些共同的经历和认知让我们这些新草地上的牧民融为一体。将这些观点整理一下，我们的任务就是，在道德的范畴下，使整个世界尽可能幸福，同时对每个人的幸福一视同仁。

但事实上，我并不会声称功利主义便是道德真理，具体来说，我不会像有些读者期待的那样，声称功利主义是经过科学证明的道德真理。相反，我想说的是，道德思维按照科学的定义对现实做出妥协和发展之后，功利主义会成为唯一可行的思想。（是否能够由此得出功利主义就是“道德真理”的结论，我不予置评。*）功利主义也许无法成为道德真理，但我相信我们依然可以使用21世纪的科学支持19世纪的道德哲学，对抗20世纪的批评观点。

下一个章节将会详细阐释，功利主义确实建立在共同价值观的基础之上。我们将从心理学、神经学以及进化论的角度对功利主义进行分析：功利主义究竟是什么？为什么说功利主义的价值观是近乎完美的通用货币？

# 第8章 通用货币出现

电影《阴阳魔界》（*The Twilight Zone*）中有一个片段极富争议性，描述了一对夫妇如何面对诱人选择的考验。一位神秘的陌生人寄来了一个小盒子，上面有一个按钮。他解释说，按下按钮，有两件事会发生：一是两人能够获得他们正急需的20万美元，二是与此同时，他们不认识的一个人会因此死去。经过一番道德挣扎和自我说服，这对夫妇按下了按钮，于是神秘的陌生人再次出现，将钱交给两人，并说——剧透慎入——现在他将提出同样的条件，把盒子送给另外一个人，一个他们“不认识的人”。

但愿我们大多数人都不会按下按钮，但毫无疑问，这是不可能的。为了描述人类共同的价值观，让我们设想一下，哪些按钮会被按下，又有哪些按钮不会被按。我们将从与道德无关的按钮开始，逐步转向与道德相关的问题。

> 问题1：幸福按钮。你会在下周遇到一条不平坦的人行道，不小心被绊倒，把膝关节摔断，你将会十分痛苦。未来的几个月中，你的幸福感也会因此骤降。但如果按下这个按钮，一点小小的魔法会使你在走路时更加留神，膝盖骨便不会被摔断了。你会按下按钮吗？当然会。我们可以得出显而易见的结论：如果其他因素保持不变，*人们倾向于让自己更加幸福。好了，下一个问题。
>
> 问题2：幸福净值按钮。这个情景中，你依然面临摔断膝盖骨的危险，也依然可以通过按下按钮来避免这场事故。但这一次按下按钮后，会有一只蚊子在你的胳膊上叮一口，让你在几天中不得不忍受恼人的瘙痒。你会按下按钮吗？当然会。以恼人的瘙痒为代价换取膝盖骨的完好是非常划算的。结论：我们都愿意衡量取舍，乐于以幸福感的小损失换取大收获。也就是说，如果其他因素保持不变，我们倾向于获得更高的幸福净值。
>
> 问题3：他人幸福按钮。这个情景与问题1十分相似，唯一的不同在于，这次的选择与道德问题有关：按下按钮不会让自己免于摔断膝盖，而是会让另一个人免受这场灾难。这次你会按下按钮吗？如果这个人是你认识并喜欢的，或者是与你同部落的一员，你当然会按；但如果你对这个人心存厌恶，也许就不会按。假设这个人是你“不认识的人”，我们姑且认为你会按下按钮。

所有人都会按下按钮吗？遗憾的是，也许不会，因为有些人是完全不在意

别人状况的精神变态。即使是正常人，面对陌生人所表现出的利他行为、冷淡，以及反感程度也是千差万别（参见第2章）。但我们要记清问题提出的背景，我们所寻找的是基于共同价值的元道德。对我们来说，共同价值并不必为每个人所认可，只要它能得到广泛认可，能得到不同部落成员的认可就好，我们希望通过共同的道德标准解决不同部落成员之间的争执。如果你极度自私，明明只需动动手指就能使他人从重度痛苦中解脱，你却依然不愿意，那么我们与你无法展开讨论。你对我们的问题不感兴趣，也不是“我们”的一员。想到这里，我们可以这样归纳：如果其他因素保持不变，人们倾向于让他人更加幸福。此外，我们还可以假定人们关心他人的幸福净值，如果你可以控制按钮，使一位陌生人被蚊子叮一下，同时免受膝盖骨折之苦，你也许依然会按动按钮。

问题4：更多人的幸福按钮。现在有两个按钮。a按钮能使一个人避免摔破膝盖，而b按钮能使10个人免于遭此厄运。你会选择a按钮还是b按钮？还是要通过掷硬币来决定？我想，你会按下b按钮。结论：如果其他因素保持不变，人们倾向于增加更多人的幸福感。

问题5：功利主义按钮。最后一个问题：a按钮能保护2个人不被蚊子叮咬，b按钮能使1个人不摔破膝盖。你会选a、b还是掷硬币？虽然a按钮能够惠及更多人，但我认为你依然会按下b按钮，因为b按钮能够避免更大的灾难。结论：如果其他因素保持不变，人们倾向于使所有人的总幸福感变得更高。

你可能会觉得这些问题很无聊，如果你确实这样想，那就太好了。提出这些问题是在为功利主义元道德的建立打基础，其答案越是显而易见，就表明基础打得越是牢固。我们已有的观点包括：首先，如果其他因素保持不变，不论是为自己还是为他人，我们都倾向于获得更高的幸福感。其次，涉及他人的考虑中，我们不仅关注个人获得的幸福感，而且关注受到影响的人数。最后，我们关注所有个体的幸福感总和，既包括每个人的幸福感，也包括受到影响的人数。如果其他因素保持不变，我们倾向于使所有人的总幸福感变得更高。

如前所述，这里的“我们”指代的并不是生活中的每一个个体，重要的是，“我们”的概念广泛、成员多样，包括来自所有部落的成员。因此我认为，世界上所有部落的成员都能够感知到上述功利主义思维的召唤，我推理的依据如下：（实验派人类学家可以走出房门，调查验证。**）如果这个观点是正确的，如果我们都能够（或经过努力后做到）在“如果其他因素保持不变”的前提下致力于增加幸福感，这就意味着我们达成了一个牢固的道德共识。这个假设在道德层面的影响是十分深远的。

当然，我们现在所拥有的只是人们会尽量扩大幸福感的假设，前面还加了“如果其他因素保持不变”的限定语。限定语的存在使上述问题显得如此无聊。假设《阴阳魔界》中那对夫妇面临的选择是：按下按钮，一位陌生人会因此死去，而另一位陌生人则会获得20万美元，那么这个片段就太无趣了，因为在本次选择中，其他因素都保持了不变。问题并不在于“我们”获得了巨大收获，而“他们”中的一员承受了更大的损失，上述问题之所以无趣，是因为获得利益者与利益受损者之间的区别不明显。因此，公平地最大化幸福感是显而易见的解决方式，同时也是十分无趣的。

现在请考虑这个选择：如果选择a按钮，会有5个人生存下来，1个人死去；如果选择b按钮，会有5个人死去，1个人生存下来。你会选择哪个按钮？如果其他因素保持不变，你当然会选择a按钮。但假设我们在讨论人行天桥困境（或者是器官移植困境），那么其他的因素已经发生了变化，随着更多细节的补充，a按钮（将人推下天桥）似乎不再是正确的选择。

这个例子说明，“如果其他因素保持不变”这个限定语十分重要。有了这个限定语，便完全不会有人对最大化幸福的努力提出质疑。但这样的结论是没有任何意义的，如果提升幸福感与我们在意的事不发生冲突，所有人都非常乐意将幸福感最大化。将这个限定语去掉，便得到了功利主义思想，这是一个完整的道德体系，是一种能够（只要有足够多的事实信息）调解任何道德争执的元道德。将限定语去掉后，我们得到的是“有意义的”结果，却失去了人们的认同，因为将幸福感最大化意味着（或者说，理论上意味着）做出道德上看似错误的事情，比如将人推向小火车。

现在我们面临的问题是：是否应当抛弃“如果其他因素保持不变”这个从句，简单地致力于将幸福感最大化？或者那样是否过于简单？我们会在第四部分解决这个问题。但首先，我们要对推崇最大化幸福感的道德思维进行更加仔细的研究，对功利主义共识背后隐藏的进化机制和认知机制一探究竟。

## 什么是功利主义？

我们知道哲学课本上的答案，但这还不够。事实上，如果把心理学和生物学比作巨大的冰山，课本上的答案不过是冰山顶上的哲学一角。

并非所有人都能被称为功利主义者，我们差得还远。但我们都能“理解”功利主义，至少从字面上看，我们都能理解为何最大化幸福感是合理的。为什么我们都能理解这种思想？为什么会存在一种系统性的道德哲学，使所有人都觉得言之有理？这种易于理解的哲学又为何总会在某件事上触动人们的道德神经，让人感到不舒服？人类大脑中似乎有一部分区域认为功利主义很有道理，而另一部分区域却会因功利主义思维而感到极端别扭，这

种情况听上去十分熟悉。

要想理解功利主义，就先要理解第2章提出的双加工体系。如果我是正确的，功利主义就像是大脑手动模式中本身存在的思想，反对功利主义的思想都是由自动模式所控制。功利主义之所以能被所有人理解，就是因为所有人大脑中的手动模式都大同小异。之所以说功利主义是唯一恰当的元道德，且为我们提供了可贵的通用货币，就是出于这个原因。

与之相对应，人类大脑的自动模式却远没有如此一致。第2章中可以看到，所有牧民大脑中自动模式的种类，即道德情感，是完全相同的，包括同理心、愤怒、厌恶、内疚、羞愧以及看到个人暴行时的不适感等。但对于不同部落、不同个体来说，触发道德情感的具体因素却大不相同。尽管触发因素不同，所有部落的自动设置中有一点是相同的：所有人的直觉反应都不可能始终与功利主义思想保持一致。因此，双加工机制的道德思维使功利主义对每个人来说都似是而非。我们都能理解功利主义，因为人类大脑拥有相同的手动模式；我们又都会因功利主义思想而感到不安，因为大脑的自动模式与功利主义思想不符。为什么会这样呢？

根据第2章（“道德机制”）的介绍，人类大脑的自动模式与功利主义不符是很正常的情况。道德思维之所以进化，是为了帮助人类传播基因，而并非为了最大限度地提升人类的整体幸福感。具体来说，道德思维的进化是为了在利己主义（“我”）和族群内部合作（“我们”）之间达到平衡，在生物学方面获得优势。而其他更像是竞争对手而非同盟的人（“他们”）则不在此考虑范围内。因此总体上讲，我们应当希望自己的道德直觉比功利主义的要求更加自私，拥有更加浓厚的部族意识。但即使大脑的进化目的是将整体幸福感最大化，高效而僵化的自动模式从本质上说也是过于迟钝的，如果它真能将我们一路引向这个目标，我们将会感到万分惊讶，更何况大脑的生物进化意义本非如此。人类大脑的自动模式不符合功利主义思想，我们也不应抱有这样的期望。

人类大脑的自动模式不符合功利主义思想，但如果我是正确的，大脑的手动模式是符合的。为什么呢？首先，让我们回顾第5章（效率、灵活与大脑的双加工机制）中关于人类大脑手动模式的概念及其产生原因的介绍，再次思考关于吃什么、何时吃的决定。概括来讲，进食对动物有利，因此存在于大脑的自动模式促使我们进食，但根据具体情况对事情进行全面考虑时，不吃有时会是更好的选择。如前所述，假如你是一名正在跟踪大型动物的猎手，不论浆果多么美味，你都没时间停下来吃。手动模式提供了灵活的空间，让我们得以忽略自动模式的倾向，（浆果！美味！）选择最有利于长期目标的行为。（大猎物！够整个村子吃一周的！）手动模式还

能促使人制订宏大的计划，看到眼前之事背后所隐藏的可能性。需要强调的是，我所说的“手动模式”并不是抽象的概念，它是一套完整的神经网络，主要由前额皮层构成，能使人类进行有意识的、可控的推理和计划。这就是我们与蜘蛛之间的区别，“手动模式”让人有能力解决复杂、新颖的问题。

“解决问题”的含义是什么？在人工智能领域，解决行为问题指的是达到某种目标状态。起初，问题求解程序有一个想法（表达法），描述了世界应有的状态，然后由程序对世界施以影响（行为），使世界变为应有的状态。恒温器就是一种简单的求解程序。恒温器包含了目标状态（目标温度）和各种影响世界的机制。恒温器非常灵活，既能加热又能冷却，还能调控加热和冷却的时间，根据温度变化调整自身行为。但作为问题求解系统，恒温器依然相对简单死板。比如，你可以将热的或者冷的物体放入恒温器的传感部分，轻易骗过恒温器的判断。

问题求解系统有很多种，但在最为抽象的层面上，它们有一些共同的特质。首先，它们都着眼于结果。目标状态就是一个人真实需要或者希望得到的结果。其次，所有的问题求解程序都会采取行动。行动的选择取决于行动与结果之间的因果关系。因此，如果上调温度能够导致预期后果（使房间达到72华氏度），恒温器便会上调温度。

将恒温器归为简单的问题求解系统，是出于如下的考虑。在某个给定的时间点，恒温器只有一个目标：使实际温度与设定温度相等。同样的，对于当前世界的状态，恒温器也只有一种“判断标准”，即当前温度。它只能做出四种不同的行为：加热、不加热、制冷、不制冷；只能“感知”两种因果关系：（1）加热会使温度升高。（2）制冷会使温度降低。最后，恒温器还能进行一些内部计算，用于判断当前温度与设定温度相比是更高、更低，还是相等。

恒温器的工作原理非常简单，但有一些有用的设施的工作原理甚至更加简单。如运动传感器，只包含一种“判断”（是否有物体在运动）和一种行为（嗨！有东西在动！）。运动传感器可以被用来解决问题（如保护博物馆的艺术作品），但其本身并不是问题求解程序。因为运动传感器并不包含设定的目标状态，也无法对世界施以影响，使其感知到目标状态已经达成。根据其感受到的情况不同，运动传感器只可能被触发或不被触发，这是一种反射型系统，属于自动模式。

毫无疑问，人类大脑的手动模式远比恒温器复杂，对于世界的当前状态，一个人会有很多不同的目标和判断，可以采取的行动也有很多，对世界上的行为和事件间的因果关系也会有很多不同的观点。但不管怎样，从最抽象的角度来看，人类解决问题的过程与恒温器的工作原理都遵从基本

的“本体论”，包括结果、行为、对世界当前状态的判断以及对世界运转情况的大致判断，也就是对可能的行为和结果之间因果关系的判断。为了理解这一原理，我们来考虑一个当下只能通过人脑解决的问题。

假设我要求你下周五中午到缅因州牛津县的邮局见我，你会因此得到一万美元。接受我的条件之前，你的大脑会对三件不同的事进行编码：一组不变的目标，你可能进行的一系列行为，一个详尽的、描述世界运行规律的模型。当我提出条件时，你的大脑便开始工作：基于你的世界模型，你推断出，如果能获得一万美元这笔意外之财，你就能更轻松地达到现有的某些目标；这个模型还告诉你，我给出的条件是可信的，遵从我的要求，你便真的能够得到一万美元。因此，你的前额皮层便把得到一万美元的价值（预计成本很低）转换成为在指定的时间和地点到达缅因州的结果。

有了新的目标后，你便要利用自己对因果关系的了解，制订行为计划。要想让你的身体移动到缅因州的牛津县，便需要一辆车，可能还需要一些机票。所以你需要打开电脑，依次按下某些按键，联系一位汽车修理工，或者向妹妹提出借车的请求。让你自己发生移动的每一个步骤都必须精心安排好，每个步骤所在的场景也都是独特的，依靠本能或自动设置的场景则不可能完成任务。因为在地球的历史上，没有任何一个生物（包括你自己）曾在你所面临的特有的交通条件下，拥有把你从家移动到缅因州牛津县邮局的尝试和经历。因此，为了完成这个任务，你需要一个完全不同的认知系统。你需要的是通用的行为计划程序，能够适应任意的目标，利用复杂的世界模型，制订出一系列详细的行为计划，最终实现目标。简而言之，这就是人类大脑前额皮层的工作。

人类大脑的手动模式为什么符合功利主义思想？我认为功利主义并不是手动模式的固有属性，事实是，一旦大脑的手动模式开始寻找道德哲学，功利主义思维便是大脑思维选择的结果。因此问题可以换成：为何手动模式倾向于选择功利主义？上文提到，手动模式的作用是达到目标状态，产生预期结果。恒温器等简单的问题求解机制并不会涉及对结果的衡量问题。对恒温器来说，唯一需要关注的就是实际温度是太热、太冷，还是刚好。重要的是，恒温器不会面临任何衡量取舍。温度不会在某些方面是好的，而在其他方面就变成了坏的。相比之下，人类的决策则充满了衡量取舍。

同恒温器一样，大脑的前额皮层也需要选择行为，达到预期结果。前额皮层会寻找路径，使你从当前状态达到预期状态，但它不会一发现行为路径便马上敦促你付诸行动。（那样的话，你就可能会为一杯柠檬茶支付600美元。）前额皮层会考虑目标的价值，也会考虑达成目标所需的成本。但这样的考虑对适应性行为来说也是远远不够的。你很渴的时候，也许愿意为一杯柠檬茶支付8美元，但如果只需2美元就能在隔壁买到相同的饮料，那么支付8美元的做法就显得很愚蠢。因此，前额皮层不仅要比较X事件的

收益和成本，还需要将X事件的净成本和净收益与Y事件的净成本和净收益进行对比。但对于适应性行为来说，这依然不够。假设花8美元买一杯柠檬茶能够免费获赠一份三明治；或者是当地的犯罪头目强烈建议你不要从隔壁那家价钱更低的店里买柠檬茶，而是从他兄弟的店里购买，这时你不仅要考虑在两家店铺里买柠檬茶造成的直接成本与收益，还要考虑连带作用的成本与收益。我们知道，真实世界中事件的结果往往不可预测，这就更增加了事件的复杂程度。

因此，通用的行为计划程序必然非常复杂，它不仅需要考虑事件结果，还要根据可能的结果及其连带作用，在选择面前进行衡量取舍。换句话说，人类大脑的手动模式是为了提供最优结果而设计的。“最优”的概念则由决策者的最终目标所决定，综合考虑行为的预期结果，包括设定的结果和可预见的连带作用。（当然，人们并不总会做出最优决策，也经常会系统性地偏离最优选择。但通常情况下，带我们走向系统性错误的往往是大脑的自动模式，而发现并辨出错误的，往往是手动模式思维。）因此，应用功利主义思维需要两个步骤，分别对应功利主义的两个基本要素。

## 从通用合理性到功利主义道德

人类大脑的手动模式由前额皮层控制，是一种通用的问题求解程序，一种优化结果的机制。但何谓最优？这个问题可以被拆为两个问题。第一，对谁而言的最优？第二，对某个给定的人来说，何种状态是谓最优？让我们从第一个问题开始入手。

假设你是完全自私的，和另外9个自私的人偶然发现了一样价值连城，但可以被替换的东西。比如装有1000枚金币的箱子，假设所有金币完全相同，你们每个人的打斗水平也完全相同，没人能够占到上风。毫无疑问，你希望把所有的金子据为己有，那么你该怎么做呢？你可以首先动武，将与你竞争的人尽可能多地打伤。但这样一来，你的竞争者便会反击，其他人也会开始动武。你可能得到一大堆金币，但也可能会一无所获，还可能会受重伤甚至死亡。

有一个显而易见的解决方案，那就是，将金币平均分给每个人，然后大家各自散去。为什么要平均分配呢？因为如果分配方案不平均，那么获得较少一份金币的人便有理由发起争端：如果有可能获得更多的金币，而且这种可能性也真实地存在了，那么为什么不为此抗争，或是声称自己将要为此抗争呢？如果群体内部的权力完全对称，那么平均分配是唯一稳定的解决方案。也就是说，在权力平衡的前提下，即使人们完全不关心“公平”的概念，所谓“公平”的资源分配方案也会自然浮现出来。功利主义的第一个基本要素——公平不倚，便由此得来。

在彼得·辛格（Peter singer）的著作《扩展的圆》（*The Expanding Circle*）中，他指出了得出公平不倚原则的另一种途径。人们并非天生公平，我们最在乎的首先是自己、家庭成员、朋友和部落内的其他成员。多数情况下，人们不会为完全陌生的人考虑太多。但与此同时，人们也认可，其他人和自己是相似的，他们最在乎的也是他们自己、家庭成员和朋友等人。最终，人们会实现认知上的一次飞跃，或者说是一组飞跃，在最高点得出结论："对我自己来说，我是特殊的。但其他人和我一样，也会认为自己是特殊的。因此，我并不是特殊的。因为即使我是特殊的，我也并不比别人更加特殊。客观地说，我的利益并不比别人的利益更加重要。"

当然，这种认知本身并不能使人信奉公平。前文提到手持金币的10位无赖也能意识到他们的地位完全对等，但他们依然是无赖。换句话说，即使人们知道偏袒自己的客观理由并不存在，也不会在主观上放弃偏袒自己的意愿。*但不知怎的，这种智力认知似乎真的被转化成了思维倾向，不论多么微弱，我们似乎对真正的公平确实产生了偏爱。我猜测这种转化是同理心，即体会他人感受能力的影响所致。人类的同理心是善变的，也是有限的，但人类理解他人感受的能力就像一粒情感的种子，有了理性的灌溉，便能开出公平不倚的道德理想之花。

说实话，我并不知道公平的理想是如何扎根在人类大脑中的。但我确信两点：第一，在我们（可以与之展开讨论的人）的观念中，公平并不是压倒一切的理想，而是我们所欣赏的一种美德。没有人完全遵照黄金法则生活，但至少我们都能理解这种观点。第二，我坚信公平不倚的道德理想是大脑手动模式的结果。当然，这种理想最初源于自动模式，源于对他人的关心。但人类的道德情感自身毫无公平性可言，只有大脑包含手动模式的生物才能领会公平不倚的理想。18世纪的亚当·斯密曾经说过，如果想到明天可能会失去自己的小指，你便会彻夜难眠；但如果知道远方的数千人明天会因地震丧生，你却依然有可能安然入睡。按照斯密的观点，我们明知道数千人丧生于地震远比我们失去一根小指糟糕，但将一根手指看得比数千名无辜的生命还要重要，这似乎是个可怕的想法。这种道德思维需要手动模式。*

你也许会问，同理心等情感是如何转化为激励式的抽象理想的？我也有同样的疑问，但无论如何，这个过程也许并不像听上去那么陌生。例如，请思考我们感到饥饿和吃饱以后购买食物的结果对比。人类获取食物的决策可能与其他动物一样，直接由自动模式驱动。吃饱以后，包括能多益巧克力酱在内的食物都完全丧失了吸引力，但我们依然可以购买食物。这种情况下，你的决定可能与饥饿时有所不同，甚至更好，不管怎样，我们确实能够完成购买食物这项任务。这是如何做到的？当然，吃饱时做出的购买

决定既受到自动模式的影响，也受到原始欲望的影响，即使已经吃饱，你依然倾向于购买你喜欢吃的，不愿买你不喜欢吃的。但同时，在吃饱的状态下购物意味着你不再直接依靠原始的食欲做出决定，相反，自动模式产生的“迫切”偏好被转化为“冷静”的认知，通过手动模式以更加实事求是的方式呈现出来。吃饱的状态下，你知道自己要为下周购买能多益巧克力酱，就像你知道塔拉哈西是佛罗里达的州府一样。

购买的决定与原始欲望之间的差别可能会以更加复杂的方式呈现。比如为别人采购时，就会考虑别人而非自己的喜好。你不仅要考虑你和别人的喜好问题，还要进行数学计算，考虑你要为多少人购买食物。同样的，你为自己采购时，也要考虑自己采购的时间频率。（我要买的是能吃一周还是能吃一年的能多益？）人类大脑似乎能以某种方式将自动模式下产生的价值观转化为动机状态，处于这种状态的思维更易受到清晰推理和定量控制的影响。这种现象的确存在，但我们还不了解其工作机制。

我们来回顾一下前文的要点：首先，大脑手动模式的本质是成本收益的推理系统，目标是得到最优的结果。其次，大脑的手动模式容易受到公平理想的影响。我认为，这种影响不具有部族差异，所有部落的成员都能理解黄金法则背后的含义。将两个观点放在一起，就能得到以创造最优结果为目的的手动模式，尽管努力的过程可能会出现瑕疵，但这种模式旨在创造公平不倚的结果，对所有人一视同仁。

现在来看第二个问题：对某个给定的人来说，何种状态是谓最优？对你、我，或是任何人来说，如何判断某个结果是好是坏？第6章（“绝妙的想法”）中，我们反复追问，“你为何看重这一点？”试着解决这个问题。例如，大多数人都重视金钱。但金钱有什么好的？你可以用钱买东西，比如能多益，比如做工精巧的小玩意。但你为什么想要这些东西呢？如前所述，如果你一路追溯，直到价值链的尽头，你便会发现自己重视的是体验的质量。广义来说，你重视的便是幸福，既包括自己的幸福，也包括他人的幸福。我说过，这个结论并不是推理的必然结果，但它无疑是自然的结论，并且是所有人都能“理解”的结论。也许并非所有价值链的尽头都是幸福，但毫无疑问的是，我们心中的很多价值链都以幸福作为终点。我们去做某件事，不过是因为喜欢；而我们会避免某件事，也不过是因为不喜欢。换句话说，我们将固有价值加到了自身的幸福和其他某些人的幸福之上，我们愿意按的按钮便可以说明这一点。虽然这种现象来源于自动模式，但它却是由手动模式造成的。我们都清楚地认为，幸福从本质上讲是价值连城的，没有人会说，“提升人们的幸福感？怎么会有人想要这样做？”

用一句话概括功利主义思想，那就是：公平不倚地将幸福感最大化。“最大化”这个概念来自大脑的手动模式，而手动模式的本质便是将事物最大化。我认为，对于所有健康的大脑，这都是通用的标准配置。通过思考“对个体而言，何者最重要”这个问题，我们得到了“幸福感”的概念。不论是你的幸福还是他人的幸福，可能都并非唯一重要，但它无疑是你发自内心所重视的诸多因素之一。我认为，这一点也是普遍适用，或是几乎普遍适用的。每个人都能“理解”幸福的重要性，稍加思考，每个人也都会发现，我们重视的很多事，甚至所有事的背后，都有着幸福的影子。最后，“公平不倚”的概念来源于某种智力认知。也许是因为人们认识到，公平的解决方案总是最稳定的；或者是同理心与人人客观平等的认知发生冲突时，人们从中获得了某种道德认知飞跃。没人能做到真正公平，但所有人都能感受到公平不倚这种道德理想的吸引力。这一点也同样是普遍适用，或是几乎普遍适用的。

因此，如果我是正确的，功利主义就是一种特殊的思想，这证明边沁和密尔在思想史上取得了前无古人的成就。他们将道德哲学从大脑自动模式的控制下夺出，抛开了人类生物学和文化历史的限制，几乎毫无保留地将其移交给了大脑的通用问题求解系统。手动模式自身并不能提供道德哲学体系，但通过植入幸福和公平这两个普遍认可的道德价值，它便拥有了创造的能力。这种结合造就了一个完整的道德体系，能够获得所有部落成员的认可。这是能够助我们走出泥潭的一条小径，一个超越道德真理冲突的体系。功利主义也许并非道德真理，但我认为，它便是我们苦苦寻找的元道德。

绝大多数的专家学者强烈反对这一观点。很多道德哲学家认为功利主义不过是来自19世纪的古怪遗迹，这种思想太过简单。尽管功利主义对道德的某些重要方面有所提及，但就其绝对性而言，功利主义将是非判断简化成为一行公式，在专家学者眼中，这是大错特错的。

## 功利主义有什么错？

我们已经遇到了有力的反驳：有时候，能够产生最优结果（以是否幸福来衡量）的行为却看似完全错误。一个经典的例子就是人行天桥困境，在这个例子中，用某人的身体来阻挡小火车能够获得更大范围的利益。

起初，我们试图通过反驳前提假设来摆脱困境：也许把人推下天桥无法达到预期效果；也许这种做法会成为影响很坏的先例等。如果你确实是这样想的，那么你确实找到了问题所在，人行天桥困境确实有些经不起推敲，不符合现实。但现在请尽量抵御这个诱惑，因为人行天桥困境的目的是解释一个更加宽泛的、需要严肃对待的问题：有时候，能够产生最优结果的

行为却看似完全错误。假设这个观点在某些情况下是正确的，那么我们能够从中得到什么启示呢？

很多道德哲学家认为，人行天桥困境突出了功利主义思维的基本缺陷。对功利主义最常见的指责就是，它低估了人类权利的重要性。很多批评者认为，用人的身体阻挡小火车这种行为本身就是错误的，即使这样做能产生更好的结果也不能改变这一事实，这个论点十分有趣，后文还会对此进行讨论。

约翰·罗尔斯是最具影响力的功利主义批评家之一，他认为在社会组织过程中，功利主义是一条差劲的原则，在考量是否将人推下天桥时，功利主义的表现也同样差劲。罗尔斯认为，边沁和密尔能够早于其他人提出反对奴隶制，应当对他们表示赞赏，但两人的反对程度依然不够。功利主义者之所以反对奴隶制，是因为他们认为奴隶制大幅削减了社会上幸福感的总和。但罗尔斯提出，如果奴隶制能够将幸福感最大化怎么办？如果是那样，奴隶制就是正确的吗？假设我们当中90%的人将其余10%的人作为奴隶，那么90%的人的幸福感就会提高，再假设拥有奴隶者获得的幸福感非常大，足以抵消被奴役者幸福感的降低。那么功利主义者似乎会支持这种粗野的不公正，就像为了拯救另外五人的生命而将一个人推下天桥一样。这个想法看上去一点都不美好。

遵循同样的逻辑，对刑事审判中性质恶劣的误判，功利主义也会表示支持。请回想第3章中法官与暴民的例子：如果阻止暴动的唯一途径是诬陷一名无辜平民并判定其有罪，那么该怎么办呢？很多人（美国人多于中国人）都认为这种行为错得离谱儿，但功利主义者可能会说，根据具体情况不同，这种做法可能是我们面临的最好选择。

情况变得更糟了。上述例子中，功利主义似乎在道德上过于松懈，允许我们随意践踏他人权利。但有些例子中，功利主义又似乎在道德上过于苛责，让我们自己的权利遭到践踏。功利主义的这些要求并不只是假设，事实上，你可能现在便面临这样的一个例子。

在你阅读的同时，世界上有上百万的人口急需食物、水源和药物；还有更多的人无法接受教育、遭受迫害而没有保护，或是缺乏政治层面的代表。富足之人视为理所应当的很多重要因素对他们来说都是奢望。例如，就在我写作的同时，乐施会美国分支（一个享有盛名的国际救援机构）正在向苏丹达尔富尔地区冲突中被困的30万名平民发放清洁的饮用水、食物、卫生设备以及其他形式的经济帮助。只需向乐施会美国分支捐赠不到100美元，就可能对这些平民的生活产生巨大的影响。我们常常听说，“只需几美元”便能挽救一个人的生命。葛武威尔是由财务分析师创立的机构，专门分析慈善机构的成本效益。葛武威尔认为，这类说法极大地低估了挽救

生命的真正（平均）成本，如果将所有成本和不确定因素都考虑在内，挽救一个人的生命大约需要2500美元。这并不是“几美元”的事情，但也完全在中产阶级的可承受范围内，随着社会的发展，也许某些较为贫困的人也会有能力承担这样的数额。

假设你每年捐献500美元，连续5年便能够拯救一个人的生命，或者也可以邀请4位好友，每人捐献500美元。再假设今年，你为自己留出了500美元的预算，不是为了购买急需的东西，而是为了自己的放松享受，比如把较为便宜的露营升级为滑雪之旅。那么为何不能把这500美元捐献给乐施会，或是对抗疟疾基金会（aMF），抑或是葛武威尔评级最佳的慈善机构呢？

需要强调的是，具体的个人选择（升级度假方式与慈善捐赠）并不重要，如果你的可支配收入不够500美元，那么50美元也可以，甚至10美元也可以。（即便你无法凭一己之力挽救一条生命，依然有很多其他善事可以做。）如果你对滑雪毫无兴趣，可以用其他非必需的小奢侈替换场景中的滑雪：比如将普通三明治升级为寿司；将虽然完好却有些旧的梳妆台换掉，购进一个设计更加时尚的新梳妆台等。同样，如果你对乐施会和对抗疟疾基金会不太熟悉，也可以将它们换掉，支持其他为有需要的人群提供帮助的慈善机构。重点在于，正在阅读本书的你很可能拥有一部分可供自由支配的收入预算，你可以将钱花在自己身上，也可以花在别人身上。这些人本身并没有做错什么，他们的需要却比你更加迫切，那么为何不把钱花在他们身上呢？

这个问题最初由彼得·辛格提出，他也是一位功利主义哲学家，继承了边沁和密尔的想法。功利主义给出的论点十分直接：将露营升级为滑雪（或其他活动）可能会提升你的幸福感，但与贫穷的非洲儿童获得清洁的饮用水、食物和住处的幸福感相比，你的幸福感实在微不足道。母亲再也不必目睹自己的孩子因饥饿或本可以治愈的疾病而死，她们的幸福感更是不必言说。因此功利主义认为，这笔钱不应用作自我享受，而应当用来帮助更需要帮助的人们。

这似乎是个好论点。事实上，我确实这样认为，第10章还会继续巩固这一论点。但认真审视这个观点，就会得出一个让人难以接受的结论：假设我们一致同意，这500美元不应花在自己身上，而是应当捐给乐施会或对抗疟疾基金会。那么如果你再有500美元呢？同样的观点仍然成立。世界上仍有很多生活无望的人，你似乎应当继续捐赠。根据功利主义观点，你需要一直不断忍痛割肉，直到将所有可支配的收入全部捐出为止。除了收入中用于提升自己能力，以便拿出更多捐款的部分之外，“可支配的”收入指的是你其余的全部收入。很显然，功利主义要求你把自己变成幸福之源，对大多数人来说，这个想法实在不好。

（抱歉地说，这并不是由功利主义思想产生的唯一糟糕想法。**）

将功利主义视为抽象概念时，即使某些道理并非显而易见，其观点听上去也是合情合理的。如果可以，怎么会有人想让世界变得更加不幸福？同样的，如果不考虑感情因素，公平地看待新草地，交战的各个部落显然应当搁置各自的意识形态，共同找出新草地上最有效的生活方式，然后按照这种方式生活。但当我们将功利主义思维应用到某个具体问题上时，它却显得如此荒谬。*在理论层面，功利主义要求我们利用人的身体阻挡小火车、恢复奴隶制、在刑事司法中造成误判，还要求我们将自己变成幸福之源。

我们究竟应当如何理解功利主义？它是我们所寻找的元道德吗？它真的是共同价值中一个合理的共识，可以用于解决道德争执吗？或者它将道德简化得过了头，已经误入歧途，如果我们拿它当真，便会走上荒谬之路？为了回答这些问题，我们需要对道德心理学进行更多了解。人类的本能反应认为，功利主义有时会错得离谱儿，这些本能反应能够代表更深层次的道德真理吗？还是它们显示了人类大脑自动模式的刻板性？换句话说，问题出在功利主义还是我们自身？新的道德认知学能够帮助我们回答这些问题。

读者们，请注意，接下来的两个章节将会十分艰涩，但为了完整表达本书观点，它们必不可缺。对功利主义的经典反驳建立在直觉基础之上，但不幸的是，要想对这些观点进行最佳驳斥，便完全不能依赖直觉。我们需要对反驳观点背后的道德机制进行科学分析，提出很多哲学论据。尽管有些读者也许并不情愿，下面两个章节会带我们深入到哲学困境假设的世界。（太遗憾了，假设问题是探索现实世界的重要工具，但人们极大地低估了假设问题的价值。*不过这已经属于另一个元哲学问题的范畴了。）

如果你认为对现代牧民来说，功利主义作为元道德已经足够优秀，那么你便可以跳过下面两章，直接进入第五部分。全书的最后两个章节中，我们将回归分歧不断的现实世界，将我们得出的结论加以应用。但在此之前，如果你想知道功利主义如何驳斥批评者们强加的罪名，便请继续读下去。大脑的手动模式，准备迎接挑战吧。

# 第四部分 道德信念

# 第9章 值得警惕的事件

在佛罗里达州杰克逊维尔市那场决定生死的高中辩论赛中，我遭遇了功利主义思想的丑陋现实：至少在理论层面，促进更大范围的利益可能意味着可怕、错误的选择，比如强制采集人体器官，比如将无辜的人推向高速行驶的小货车，等等。难怪包括思想深刻的哲学家在内的很多人都认为，进行是非判断时，最大化幸福感这个标准远远不够。

接下来的两章中，我们将直面挑战。总体策略共有两个，一是协调，二是改革。通过协调，我将证明将幸福感最大化并不会得出显而易见的荒唐结论。也就是说，功利主义为现实问题提出的解决方案大体上与常识吻合。例如，我们能以很多方式证明，即使初衷再好，将人推下天桥和偷取人类器官在现实世界中都不可能长期促进更大范围的利益。*

然而，功利主义不可能只是关于常识的哲学主张，因为全世界的道德部落都有各自不同的常识系统，这也是常识道德悲剧发生的原因。从功利主义的角度来看，全世界的部族道德不可能同样美好，这就意味着功利主义思想必然会与其中一些常识产生矛盾。功利主义最初在19世纪的英国产生，作为社会改革的理论基础出现。要求改革意味着挑战传统智慧，要想有效挑战传统智慧，改革者就需要阐明，尽管传统智慧有诸多好处，但它是不正确的。例如，密尔曾挑战当时的传统智慧，认为女性的智力水平与男性相当，具体来说，他认为当时女性的聪明才智不及男性是因为她们无法接受教育。

密尔采用的是揭露式的论证方法，揭示为何本身错误的事情会看似正确。这两章所采用的改革策略也是这样，揭示看似正确的道德真理为何错误。具体来说，我们会利用科学，对反对功利主义的道德直觉进行研究；理解道德直觉的作用；解释为何过于刻板的道德直觉不能作为是非对错的最终评定标准。*

人类大脑的自动模式，即道德直觉，可能在两个方面让我们失望。一方面，道德直觉可能会过于敏感，稍加思索便会发现，有些事件与道德的关系没有看上去那么大。例如，研究表明，死刑案件的陪审团常常对被告的种族背景十分敏感，但在我们（可以与之展开讨论的人）今天看来，种族背景与道德其实并不相干。另一方面，自动模式也可能会过于不敏感。稍加思索后，有些事情确实事关道德，但自动模式却没能做出相应的反应。例如，陪审团提供审判意见时，可能并未认真考虑被告在实施犯罪时的年龄，但今天的我们认为，年龄确实是事关道德的因素。

下面我们将从上述两个方面提出证据，说明反对功利主义的道德直觉为何不可靠。首先，我们从最受喜爱的道德果蝇—小火车困境开始，随后将会证明，人类对假设案例的直觉反应与真实世界的问题的确相关。（参见第4章关于医生和公共卫生专家不同道德判断的研究，这部分内容已经部分证明了这一点。）

## 按下道德按钮

先来回顾小火车问题的一些基本事实：面对开关困境，多数人都认为应当扳动开关，让小火车避开5个人，撞向1个人。面对人行天桥困境，多数人反对将人推下天桥，挡在小火车的轨道上，从而以一命换五命。这是一个心理学问题：同样都是一命换五命，为什么我们能接受开关困境中的做法，却无法接受人行天桥困境中的做法？

回溯第4章，这个问题已经得到了部分解答。针对将人推下天桥的行为，情感会自动做出抵触反应，但针对扳动开关的想法，却不会产生类似的情感反应。两种情况都采用了功利主义成本收益分析的思维（“牺牲一人来拯救五条生命是更好的选择”）。但只有在人行天桥困境中，典型的情绪反应非常强烈，胜过了功利主义思维过程（参见图4.3）。

前文已经提到，很多研究结果都与双加工体系的解释相符，研究的手段具体包括功能性脑成像技术、观察有情感缺陷的神经病患者、对情绪触发进行心理学衡量、情绪诱导、人为扰乱手动模式的思考过程（时间限制、干扰性任务）、人为干扰视觉表象、性格调查问卷、认知测试、药理学介入等。但双加工体系的解释并不完整，它无法解释为何人行天桥困境比开关困境牵涉的情感更多。人行天桥困境的何种因素触动了大脑的情感开关？

探寻正确答案之前，我们需要首先回顾之前的证据，将小火车问题中一个诱人的错误答案排除掉。如前所述，开关困境中，拯救五条生命的行为是可能实现的，但在人行天桥困境中，可能出错的环节有一百万个。一个人的身体真的能够挡住小火车吗？如果这个人没有落到轨道上怎么办？如果这个人进行反抗怎么办？诸如此类的问题还有很多。也就是说，从功利主义思维的角度看，现实世界中扳动开关的理由很多且充分，但将人推下天桥则完全不同。毫无疑问，这种考量是正确的，但证据表明，这并非人们拒绝后者的原因。如果人们的否定选择真的是基于现实考虑，基于实际的成本收益分析，那么为什么在认知反应测试中得分较高的人们做出否定选择的概率会更低呢？在时间有限的情况下，为什么人们做出否定选择的概率又会升高？为什么有情感缺陷的人们和视力不佳的人们做出否定选择的概率会降低？如此等等。这些结果表明，拒绝将人推下天桥这个决定出自直觉反应，而并非出自超现实的成本收益计算。（后文描述的试验对人们

在现实世界的期望进行了控制，还会为此提供更多的证据。**）如果让我们拒绝将人推下天桥的是直觉反应，而不是现实的功利主义思维，那么这种直觉反应是由何而触发的呢？

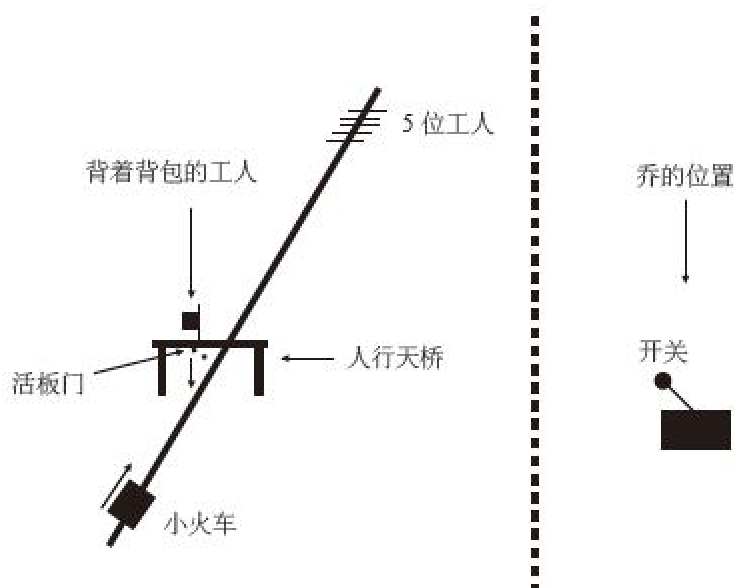



图9.1 远距离人行天桥困境

第4章中，我将类似人行天桥困境的案例称为“主观的”，将类似开关困境的案例称为“客观的”。对于不同版本的人行天桥困境来说，有害行为的“主观”程度不同，通过对比这些版本，我们便可以验证一个结论。*首先，我们将原始的人行天桥困境中造成伤害的触发方式换为扳动远处的一个开关，就像开关困境一样。这个版本叫作远距离人行天桥困境。

远距离人行天桥困境中，我们的主角名叫乔，他可以扳动开关，打开天桥上的活板门，使工人掉下天桥，落在轨道上，挡住小火车，挽救5个人的生命（参见图9.1）。原始的人行天桥困境测试中，31%的参试者支持将人推下天桥，以挽救另外5人。随后，我们向另外一组情况完全相同的参试者提供了远距离人行天桥困境，表示支持的参试者比例达到63%，几乎是此前的两倍。试验表明，“主观的”因素确实会影响人们的判断。

远距离人行天桥困境与原始版本的不同之处在于，动作的执行者距离受害者较远，而且动作的执行者不会触碰到受害者。那么问题的关键是距离？是触碰？或是两者兼有？为了弄清这一点，我们设计了人行天桥开关困境

（参见图9.2）。这个场景与远距离人行天桥困境相似，只是开关的位置设在了天桥上，就在受害者旁边。

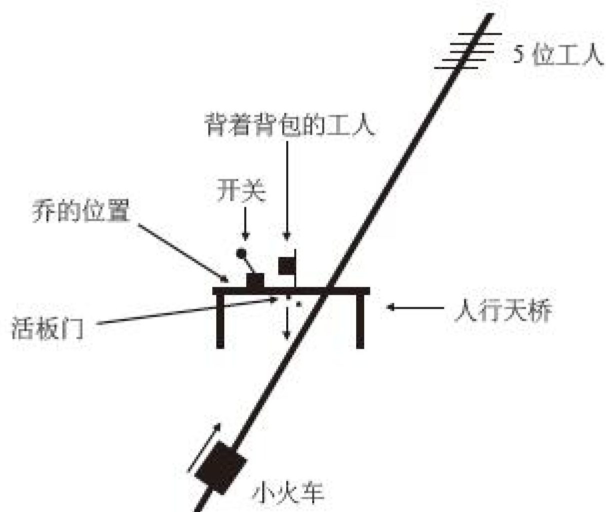



图9.2 人行天桥开关困境

这一次，59%的参试者对功利主义行为表示支持，与远距离人行天桥困境的试验结果十分接近，在统计学上基本没有区分意义。因此，空间上的距离对人们的选择似乎没有影响，或者影响很小，关键因素也许在于是否发生触碰。*

即使是现在，我们依然面临多种解释。人行天桥困境中，动作的执行者触碰到了受害者，人行天桥开关困境中则没有发生触碰。但人行天桥困境中动作的执行者做出了更加微妙的动作，他通过自己肌肉的力量直接对受害者施加了影响：他推了受害者。这就是个人力量的应用。为了区分触碰和运用个人力量，我们设计了人行天桥长杆困境（参见图9.3）。这个场景与原始的人行天桥困境相似，只是这个场景中，动作的执行者需要通过长杆推动受害者。这样便在不发生触碰的情况下加入了个人力量的影响。*

这一次，33%的参试者支持将人推下天桥——数量明显下降。与远距离人行天桥困境和人行天桥开关困境相比，这一数字几乎减半。此外，表示支持的人数比例33%与原始人行天桥困境测试中得到的31%并没有统计学差异。

因此，人行天桥困境与开关困境之间重要的心理学差异是造成伤害的“主

观”程度，具体来讲，则是个人力量的应用——是亲手推还是扳动开关。

从规范的角度思考，这个结论十分有趣，因为我们根本不会认为个人力量与道德判断有关。对某人的性格进行评判时，是否愿意利用个人力量对他人造成伤害确实是相关的考虑，也就是说，如果某人愿意亲手杀死别人，你便有充足的理由认为这个人十分糟糕，相比之下，以更加间接的形式造成伤害（协调*）则会稍好一些，但在这个过程中，个人力量的介入并不会使这种行为更加糟糕（改革）。请这样想：假设一位朋友在人行天桥上给你打电话，寻求道德建议：“我应当为了拯救五条生命而杀死一个人吗？”你肯定不会说，“嗯，这需要看具体情况……你会直接推那个人吗？还是你可以通过开关让他掉下去？”显而易见，触碰这种物理机制本身与道德并不相干，但却与心理学的考虑密切相关。

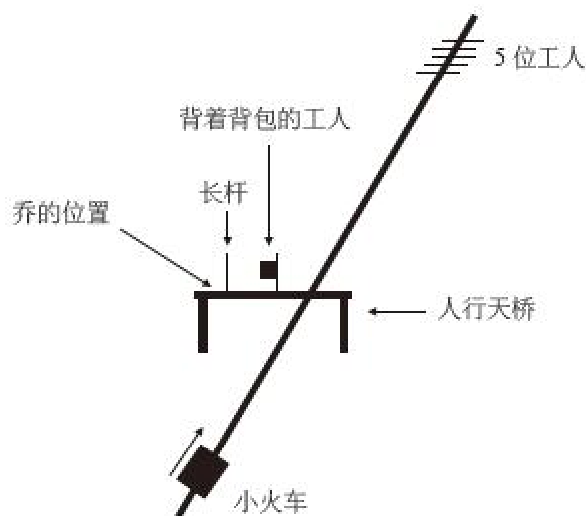



图9.3 人行天桥长杆困境

这正是双加工理论所预见的情况。根据第4章的所有科学描述，我们了解到，自动模式让我们在人行天桥困境中拒绝将人推下天桥，这是直觉的情感反应。我们从第5章得知，自动模式是刻板的探索式机制，在有些情况下并不可靠。但这种不可靠体现在哪些方面呢？

根据目前掌握的情况，我们无法给出确切答案。大脑的自动模式也许不够灵敏：让人从活板门中掉下去也许确实不对，但因为过程中没有包含“推”的动作，我们便没有提起足够的警惕。自动模式也可能过于灵敏：“推”这个动作是正确的，与“不推”导致的五倍于此的伤害相比，自动

模式对受害人所承受的损害有些过于担心。稍后我们会继续讨论这个问题，目前得到的结论是：从某个角度来看，大脑的自动模式正在将我们引入歧途。

## 手段与连带作用

我曾在第4章中提出暗示，开关困境和人行天桥困境之间还有一点重要区别：造成伤害是达成目标所需的手段还是过程中产生的连带作用。在人行天桥困境中，我们要使用某人的身体阻挡小火车，而在开关困境中，受害者的死亡则是连带作用的结果，是“附带损伤”。为了理解两者的区别，可以假设受害者突然神奇消失，然后比较两者的结果。人行天桥困境中，受害者一旦消失，整个计划便会失败，因为小火车无法被阻挡；但在开关困境中，支线铁路上工人的消失则会成为上帝的福音。

哲学历史上，手段与连带作用的区别由来已久，甚至可以追溯到圣托马斯·阿奎那（公元1225—1274年）的时代。他创立了“双重效应原则”，其本质就是“连带作用原则”。双重效应原则认为，如果为了达到目的所采取的手段造成了伤害，这种做法是不对的；但如果追求好的结果时产生了连带作用，对某人造成了伤害，则是可以容许的。同样的，第4章中提到，康德认为道德法律要求我们将他人“在任何时候都看作目的，永远不能只看作手段”。

手段与连带作用的区分在现实世界中十分重要，不论是刑法、生命伦理学还是国际战争法则。例如，“战略轰炸”和“恐怖轰炸”的区分便是以手段与连带作用的区分作为基础。如果轰炸平民的目的是为了打击敌方士气，便是恐怖轰炸，为国际法所禁止；但如果轰炸目标是兵工厂，导致兵工厂周围的平民丧生，造成“附带损伤”，这种轰炸就属于战略轰炸，国际法中并没有明文禁止。同样，美国医学会认为，以终结慢性病人生命为目的有意施用大剂量止痛药的行为须被严格禁止；但怀着减轻病人痛苦的目的，施用足以终结病人生命的过量止痛药则属于未尝不可的行为。

大脑的自动模式对手段与连带作用的区分敏感吗？人们对不同的小火车困境做出的不同反应是否能从这个角度得到解释？为了弄清这一点，我们将原始的人行天桥困境与另一个相似场景进行比较，这个场景中，伤害以连带作用的形式造成。请考虑障碍物冲撞困境：

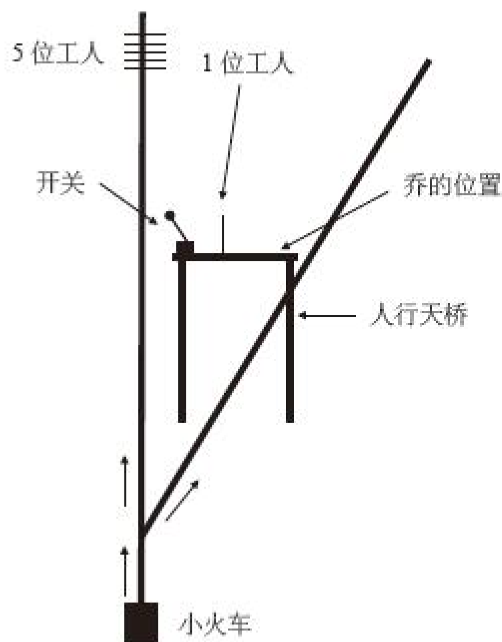



图9.4 障碍物冲撞困境

小火车冲向了5位工人，旁边有一条支线，上面有一位工人。与开关困境相同，我们可以扳动开关，使小火车开上支线，挽救5个人的生命。故事的主角乔站在横跨支线铁路的一架人行天桥上，天桥又高又窄，能让小火车改线的开关位于天桥的另外一头。糟糕的是，在乔和开关中间有一位工人，为了挽救5个人的生命，乔必须迅速扳动开关，因此，他必须尽可能快地跑过天桥，但这样就会将那位工人撞下天桥摔死。与原始的人行天桥困境相同，这个场景中，造成伤害的行为也属于个人行为，因为乔是通过个人力量将工人撞下天桥的。但不同的是，这一次的伤害由连带作用导致，属于附带损伤。如果这位工人能够神奇地消失，那便是皆大欢喜的好事。

这一次，尽管知道乔的行为会产生连带作用，导致工人死亡，仍有81%的参试者对挽救5人生命的行为表示支持。这是相当高的支持率，远远高出原始版本的31%。此外，这个数字与原始开关困境中87%的支持率也相去不远，在统计学上并没有区分意义。因此，大脑的自动模式似乎对手段与连带作用的区分高度敏感。这就解释了为何人们会在开关困境中给出肯定答案，而在人行天桥困境中给出否定答案。

分析到这里，我们看似是在为自动模式辩护。我们认为，手段与连带作用

的区分与道德问题紧密相关。人类的道德直觉似乎也盯紧了这一点，当受害者因达成目标所需的手段而受到伤害时（人行天桥困境，人行天桥长杆困境），便给出否定答案；当受害者因连带作用受到附带损伤时（开关困境，障碍物冲撞困境），便给出肯定的答案。但还有一个小故障：远距离人行天桥困境和人行天桥开关困境中，都使用人的身体来阻挡小火车，但人们大多选择了肯定的答案，支持率达到60%。情况变得更加棘手了。

哲学小火车问题出现早期，朱迪斯·贾维斯·汤姆逊（Judith Jarvis Thomson）提出了这样一个场景，我们将其称为环线铁路困境。这个场景与开关困境类似，但这里的支线铁路与干线相通，如图9.5所示。

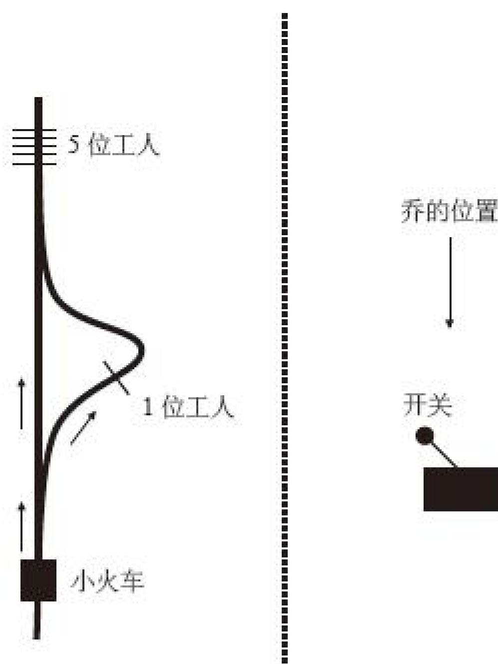



图9.5 环线铁路困境

这个场景中，如果支线铁路上没有人，小火车便会回到干线铁路，撞上5位工人。也就是说，这个场景中扳动开关就相当于利用受害人的身体阻挡小火车，挽救5条生命。（如果支线铁路上没有人，扳动开关没有任何意义。*）尽管如此，仍有81%的参试者支持扳动开关。因此在有些情况下，利用人的身体阻挡小火车似乎是合乎道德的。

此外，还有一个场景也会干扰手段与连带效应的区分。（坚决拥护小火车

问题结论的人们，请注意，这个场景对“三重效应原则”也会造成干扰。）我们将这个场景称为碰撞预警困境。造成伤害的方式与原始开关困境完全相同，但这一次受害者被当作达成目的的手段。（参见图9.6）

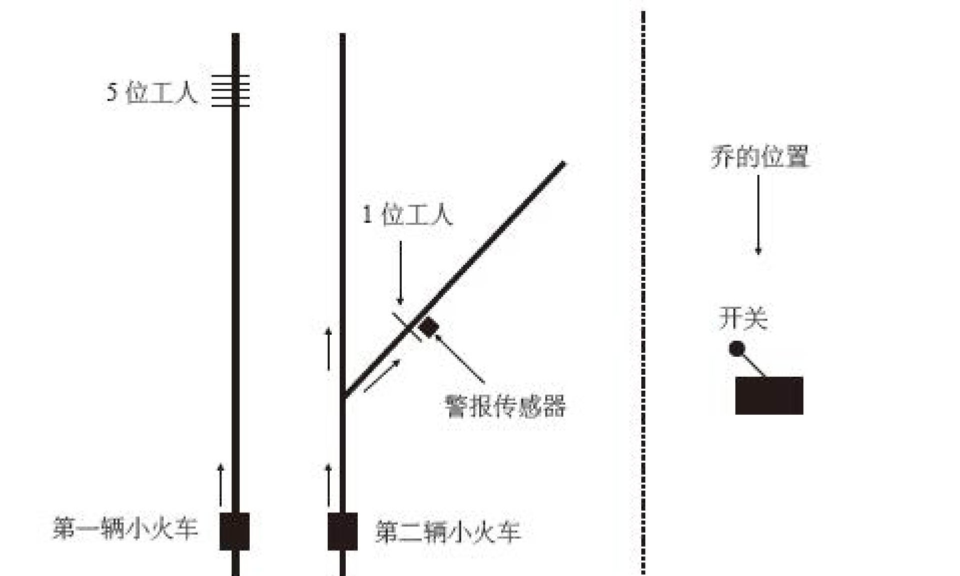



图9.6 碰撞预警困境

整个场景是这样的：第一辆小火车冲向了5位工人，如果不采取任何措施，5人都将死去。第二辆小火车行驶在另一条铁路上，前方没有任何障碍。乔可以扳动开关，使第二辆小火车开上支线铁路。支线铁路上有一位工人，身边有一个与警报系统相连的传感器。如果乔扳动开关，第二辆小火车就会开上支线铁路撞倒工人，传感器会感应到这起事故，触发警报，整个火车系统会因此断电，使第一辆小火车失去动力，挽救铁路上的5条生命。这里的关键因素是：撞倒受害者是挽救另外5人的手段。

这个场景中，86%的参试者支持符合功利主义思想的行为，与原始开关困境中87%的支持率不相上下（统计学上没有区分意义）。在这个场景中，尽管受害者作为挽救他人的手段被牺牲，但人们依然支持一命换五命的决定。

这是怎么回事？小火车问题中，我们已经找到了两个影响人们直觉判断的因素：伤害是否由个人力量直接造成（推的动作与扳动开关），以及受害者是因达成目标所需采取的手段受到损害还是因连带效应而受到损害（利用人体阻挡小火车与附带损伤）。但这两个因素造成的影响并不稳定。个

人力量有时十分重要，比如人行天桥开关困境与人行天桥长杆困境的对比；有时却并不重要，比如开关困境与障碍物冲撞困境的对比。同样的，手段与连带作用的区分有时十分重要，比如障碍物冲撞困境与人行天桥困境的对比；但有时却并不重要，比如开关困境、环线铁路困境和碰撞预警困境的对比。这些因素为什么有时重要有时不重要呢？

仔细思考便会发现，只有将两个因素组合起来，才会产生重要的影响。使用个人力量对某人造成伤害时，如果伤害只是连带作用的结果，这种行为便是可以被接受的（障碍物冲撞困境，支持率81%）。如果为了达到目的而对某人造成了伤害，但其中没有使用个人力量，这种行为也是可以接受的（环线铁路困境，支持率81%；碰撞预警困境，支持率86%）。但如果为了达到目的而对某人造成了伤害，而且在其过程中使用了个人力量，那么大多数人都不会认可这种行为（人行天桥困境，支持率31%；人行天桥长杆困境，支持率33%）。因此，将伤害作为达成目标的手段，并使用个人力量加以实现，两者似乎是一个神奇的组合。**（这种现象被称为“相互作用”，与药物的相互作用类似：同时服用两种药物产生的效果优于单独服用两种药物产生的效果之和。）

我们已经围绕小火车问题设立了足够多的场景，现在让我们稍作休息，对这些结论的重要意义进行反思。在这些场景中，我们试图弄清道德直觉究竟是否可靠。但结论如何呢？似乎并不太可靠。一方面，我们的判断有时会对个人力量的介入（推的动作与扳动开关）十分敏感，但这个因素似乎与道德关系不大。大脑似乎将手段与连带作用的区别作为某种判断依据，但其判断过程却并不严谨。尽管人们大多偏爱手段与连带作用的因素，嫌弃个人力量的介入，但我们发现，大脑对两种因素的敏感程度紧密相关。*

还有另外一个重要问题。显而易见，推和扳动开关两个动作本身的区别与道德无关。但倘若你以此询问一位德高望重的道德哲学家，她会告诉你，因达成目标所需手段造成的伤害与连带作用导致的伤害在道德上是不同的。为什么呢？在充分了解后果的前提下，以附带损伤的形式导致某人死亡为什么会比以杀人为目标导致某人死亡好呢？不管怎样，即使你因附带损伤而死，你终究还是死了，而杀死你的人也早就知道你会因他而死。（请注意，这里讨论的是可预见的连带作用，就像开关困境中的情况一样，意外情况暂且不列在讨论范围内。）长久以来，人们都认为以杀人作为达成目标的手段比连带作用致死更加糟糕，这一观点也得到了广泛的认同。但据我所知，备受推崇的双重效应原则并没有实证支持，勉强支撑这一原则的只有一些直觉感受，人们对此也是心知肚明。世界各地的人们都将双重效应原则作为依据（参考），进行道德判断，但他们事实上并不了解这个原则。这说明，人们将直觉判断排在第一位，双重效应原则不过是

对直觉判断的（不完整）归纳。也就是说，因达成目标所需手段造成的伤害与连带作用导致的伤害之间存在“基于原则的”区别，我们不能由双重效应原则推出直觉判断的合理性，恰恰相反，这项原则的合理性应当由道德直觉是否合理决定。

那么，直觉判断从何而来？与因可预见的连带作用造成的伤害相比，为什么因达成目标所需手段造成的伤害往往会给人更差的感觉呢？下文当中，我将提出一个理论，阐明为何人类的道德思维对手段与连带作用的区别如此敏感。如果这个理论是正确的，那么我们手中的道德遗产——一度被奉为真理的双重效应原则将会遭受极大的质疑。真实世界中，很多政策的制定都以双重效应原则为基础，这些政策影响着人们每天的生生死死。

## 模块近视

我将这个理论称为模块近视假设，它将道德判断的双加工理论与思维对行为的表征理论相结合。本书介绍的所有观点中，模块近视假设是最为复杂的一个，没有之一。为了便于理解，我将首先对本理论进行总结，提出概括性观点，然后在接下来的部分逐步展开论述。

模块近视假设的基本概念是这样的：首先，人类大脑包含一个认知子系统，也就是一个“模块”，其作用是监控个体的行为计划，并在个体打算伤害他人时敲响情感的警钟。**其次，这个报警系统是“近视眼”，无法识别有害的连带作用。这个模块会对行为计划进行审查，排除伤害。但出于某些原因（稍后解释），这个审查模块无法“看到”因连带作用而产生的伤害。只有人们计划以伤害他人作为达成目标的手段时，才会被模块发现。认知子系统模块肩负发出警示的责任，阻止人类做出基本暴力行为。模块近视假设指出了认知子系统的局限性，解释了为何直觉总要将手段与连带作用区分开来。认知子系统的局限性使得某些伤害成为人类的情感盲区，但并非认知盲区，也就是说，人类在情感上也许没有反应，但对该事件的认知不会受到影响，这个概念也许有些耳熟。稍后我会解释，这个概念的二重性反映的就是道德判断中的双加工理论。

这就是模块假设理论的概括性描述，这个描述引出了两个重要的问题：首先，人类大脑为何要设立一个系统，专门对行为计划进行审查、排除伤害？下文将会解释这个机制对我们有何意义。其次，为什么近视模块会出现这样的盲区？稍后我将说明，近视模块假设的完美之处在于，它由道德判断的双加工模式自然发展而来，同时融入了描述大脑如何呈现行为计划的一项理论。

## 我们为什么没有成为精神变态？

人类大脑为何需要行为计划审查系统？我的假设是这样的。

在人类发展史的某个阶段，我们的祖先获得了对复杂行为进行计划的能力，他们能够思考长期目标，并想出创造性的解决方案。也就是说，我们的祖先获得了手动模式控制下的推理和计划能力。这是一个了不起的进步，他们由此能够进行有组织的狩猎并布置陷阱，猎杀大型动物；能够建造结构更加合理的房屋；能够播下种子，期待几个月后的收获等。但整体来看，对长期目标进行思考并设计解决方案的能力会导致一个可怕的副产品：预谋暴力的大门被打开了。暴力行为的来源不再局限于当下的冲动，而是可能作为一种多功能工具，帮助人们获得自己想要的东西。不想再听命于那个蠢人了？伺机除掉他吧！喜欢那位邻家姑娘吗？等到她孤身一人的时候想办法搞定她吧！能够制订未来计划，想出新型解决方案的物种是非常危险的，特别是如果这个物种还会使用工具，危险便又加深了一层。

对黑猩猩来说，想要杀掉另外一个同类并非易事，如果对方体型更大，更加强壮，则更是难上加难。霍布斯曾说过，对于人类这样的灵长类动物来说，每个健康的成年个体都能独自杀死种群内部的任意一名成员，这个事实十分有趣，可也令人感到毛骨悚然。一名3.5英尺高的女人可以趁身高6.5英尺的男人睡觉时偷偷上前，用一块石头砸碎他的脑袋。因此，当人类能够熟练地制订行为计划并使用有利于自己的工具时，我们使用暴力的能力也得到了极大的提高。

使用暴力的能力提高有什么坏处呢？对于老虎等画地而居、习惯独处的动物来说，也许没有问题。但人类生存的基础是组成合作群体，共同生活，遭到攻击的人们往往会实施报复（“以牙还牙”），暴力行为对攻击者和被攻击者都十分危险。尤其是当某人的目标受害者会使用工具，并能够计划自己的行为时，即使攻击者拥有两倍于受害者的体型，只要受害者能够活下来，他便可以寻找机会实施报复，可以用石头偷袭头部或者用匕首刺穿后背。即使受害者没能活下来，他的亲戚或朋友依然可能会以他的名义积极实施报复。在有仇必报的世界里，如果每个人都有能力杀死另外一个人，我们对待他人的态度便要加倍小心。更重要的是，抗拒暴力的心理也许不能为单独的个体带来任何优势，但在群体的层面上，这种心理确实是一种优势。（群体内部）态度温和的群体更愿意合作，也因而能够获得更大的生存优势。简而言之，滥用暴力的个体可能会遭到群体成员的报复，会影响整个群体合作的能力，进而将整个群体置于群体竞争的不利境地。

为了控制暴力行为，我们需要某种形式的内部监控，只要使用暴力的想法一出现，就会有报警系统提醒人们“不要这样做”。这种行为计划审查机制无须反对所有暴力，当人们需要自卫或攻击敌人时，这种机制便无须启动。但大体来讲，这种机制需要使个体对相互之间的身体伤害十分抵触，从而降低报复造成的伤害，支持群体内部合作。我的假设是，近视模块就

是人类大脑的行为计划审查机制，负责阻止人类滥用暴力。

那么为什么这个模块是“近视的”呢？因为从某种程度上说，一切模块都是近视的。我们假设了一个小型报警系统，用于监控以结果为导向的手动模式，对其制订的计划进行审查，排除潜在危险。（也许你会想起，由于涉及个人得失，这种机制远非公平不倚。）所有的自动模式都是探索性的，因此必然会在某些方面无法明察秋毫。例如，由杏仁核控制的认知系统能通过放大的眼白识别恐惧。但该系统无法判断触发系统反应的“眼白”究竟是电脑屏幕上的图像，还是真实的人类身处危险境地时睁大的眼睛。为了实现识别目标，所有的自动模式都需要依靠某些特定的线索，但这些线索往往并非完全可靠。同样的，我们设想的反暴力报警系统不论多么有效，都必须基于某些特定线索做出回应。因此，我们并不必考虑自动的行为计划审查机制是否近视，而应当问：这种机制的盲区是哪个方面？

至此为止，我们的讨论仅仅停留在理论层面，是否真有证据表明此类机制的存在呢？答案是肯定的。我们已经知道，对于将人推下天桥等形式的暴力行为，我们会自动产生情感反应。我们也知道，这种机制在某种程度上是“模块化”的，也就是说，其内部活动与大脑其他部分的活动互不相扰，至少与负责有意识思维的大脑区域相互独立。（我们为何无法通过单纯的思考判断小火车困境中直觉的反应，而必须借助之前描述的试验得出结论，原因便在于此。）基于试验的小火车理论表明，大脑中确实存在某种反暴力的自动机制。

基于模块近视假设，可以进一步做出三个预测。第一，该模块最初并不是为解决小火车困境而产生，真正的触发因素应当是现实世界中的暴力行为。第二，如果该模块的触发因素是与暴力有关的线索，那么并不必发生真实的暴力行为，只要有相应的线索，系统便会被触发。也就是说，以正确的方式模拟暴力行为便足以触发模块反应。即使在手动模式层面，参试者明知不会有真正的暴力行为发生，只会有模拟行为，模块也会发生反应。第三，如果报警系统的作用是阻止无端暴力行径，那么在亲自模拟暴力行为、观看他人模拟暴力行为和模拟姿势相仿的非暴力行为三种情况中，系统应当对亲自模拟暴力行为做出最为强烈的反应。

基于这些预测，菲耶里·库什曼、温迪·门德斯和同事们进行了一项试验，第2章已有介绍。库什曼和同事们让参试者模拟用锤子击打他人腿部、击打桌子上婴儿的头部等暴力行为（参见图2.2）。人们亲手模拟暴力行为时，外周血管会剧烈收缩，产生“胆战心惊”、手脚冰凉的现象。但参试者观看他人模拟暴力行为，或者亲自做出姿势相仿的非暴力行为时，则不会产生类似反应。试验过程中，参试者（在手动模式层面）完全清楚其行为不会造成任何伤害，但试验者依然观察到了上述各种反应。因此，库什曼和同事所观察到的现象恰好与模块近视理论的预测相符：亲自模拟与暴力

行为相似的行为时，人们会自动产生抵触心理。

然而，近视模块假设还不止于此。人类大脑拥有报警系统，能够对与暴力相关的线索做出反应。根据我们的假设，该系统有一个特定的盲区，它无法识别因可预见的连带作用造成的伤害。这又是为什么呢？

## 无法识别连带作用

事情开始变得复杂了。阿尔文·戈德曼（alvin Goldman）与迈克尔·布拉特曼（Michael bratman）最早提出相关概念，约翰·米哈伊尔（John Mikhail）随后对其想法进行扩展，提出了行为表征理论，我们所提出的理论便从此开始。米哈伊尔的观点是，对行为计划进行逐步分析时，大脑对行为的表征方式如图9.7所示。图9.7展示了开关困境和人行天桥困境中动作执行者的行为计划。

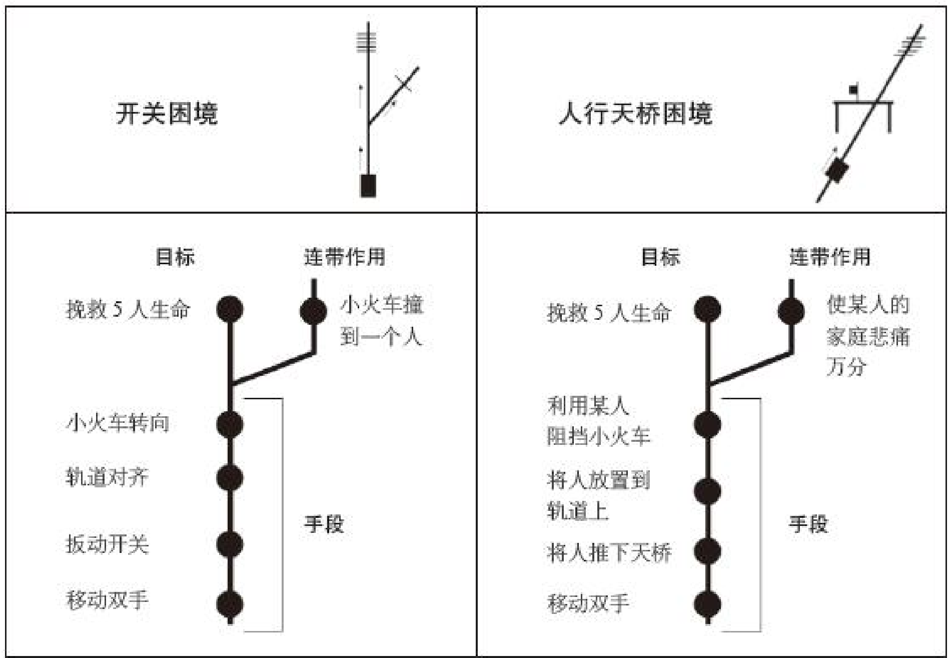



图9.7 开关困境和人行天桥困境的行为计划

每个行为计划都包含一条最主要的线索，即“主线”，开始于执行者的肢体动作，结束于执行者的最终目标（预期结果）。主线列出了达成目标所必需的、环环相扣的每个步骤。例如开关困境中，动作执行者只有移动双手（肢体动作）才能切换开关；然后铁轨上的转辙器才会移动，对齐支线铁

轨；小火车才会开上支线铁路，而不是在主线铁路上继续行驶；主线铁路上的5个人才能得救（最终目标）。同样的，人行天桥困境中，动作执行者只有移动双手（肢体动作）才能导致工人摔下天桥；然后工人会落在铁轨上；小火车会撞上这个人；然后小火车会停止；5个人才能得救（最终目标）。图9.7分别按顺序标出了重要事件，可以从两条主线的底端开始向上回溯。行为计划中，主线上列出的事件是达成目标所必需的手段，是目标达成过程中的必要步骤。

图9.7的行为计划图中还包括次级线索，即主线上的分支。开关困境中，小火车的转向产生了两个效果（“双重效果”）。其一是导致主线铁路上的5个人得救（目标），同时也产生了严重的连带作用：支线铁路上的1位工人因此死去。这个事件属于可预见的连带反应，因此被标注在了分支线索上。动作的执行者能够预见到这个事件的发生，但该事件并非达成目标的必需环节。（如前所述，如果支线上的工人突然消失，事件的目标依然能够达成。）同样的，人行天桥困境中也包含可以预见的连带作用。利用人的身体阻挡小火车能够挽救5条生命，但不难想到的是，该行为还会导致其他后果，比如受害者的家人会因此悲痛万分等。图9.7将这个事件标注在了分支线索上，这是可以预见到的结果，但并非达成目标的必需环节。即使受害者的家庭对这个结果十分满意，整个行为计划也不会受到影响。

米哈伊尔的理论相对直接：我们认为将人推下天桥的做法不对，是因为这个人被当作达成目标的手段。我们认为扳动开关，使小火车转向的做法可行，是因为这个人的死亡是由于可预见的连带作用导致的。米哈伊尔的观点是，某种形式的心理表征，即以不对称的分支结构所呈现的行为计划，可以作为自然的模型，展现手段与连带作用的区别。这个观点十分简练。

第一次看到米哈伊尔的理论时，我认为他的观点十分有趣，但却不敢苟同。首先，支持双加工理论的证据已有很多，这些证据都表明，由大脑中某个区域控制的情感反应会与另一个区域控制的功利主义判断相矛盾。但米哈伊尔的理论并未涉及情感问题，也不存在不同系统之间的矛盾，只有一个单独的系统，一个“通用道德语法”系统，对分支行为计划进行无感情的呈现和分析，从而完成所有判断。因此，不论米哈伊尔的理论多么诱人，他的基本研究方向就是错误的。第二，更有说服力的是，手段与连带作用的区别无法对相关数据做出合理解释。起初，朱迪斯·汤姆逊提出的环线铁路困境对手段与连带作用理论起到了一锤定音的作用。环线铁路困境在本章的前一部分已有介绍，这个场景同样涉及利用人的身体阻挡小火车，但人们对此似乎并没有过多异议。

后来，在一个酷热难耐的夏日，我站在费城老屋的空调窗机前乘凉，突然之间，我感到眼前灵光一闪。其他一些研究者的试验数据表明，手段与连带作用的区别确实适用于各种类型的事例。相较之下，人们对包含负面连

带作用的事例更加支持，而对将伤害作为手段达成目标的事例则更加反对。（将人行天桥困境与障碍物冲撞困境比较时，也会得出同样的结论。）这些研究表明，手段与连带作用的分别确实有意义。

如何才能将这些结论与双加工理论统一起来？前文提到，双加工理论认为，将人推下天桥等令我们不快的行为会触发负面的情绪反应。因此我想，如果手段与连带作用的区分确实有意义，那么这种区分一定对触发情绪反应的机制造成了影响。也就是说，对于将伤害作为手段达成目标的事例，我们的情绪反应一定会更加激烈。但环线铁路困境明显涉及利用受害者阻挡小火车，我们为何会对此淡然处之？是因为环线铁路困境有什么特殊之处吗？这个场景是否包含了某些因素，使我们不再对伤害感到不安？想到这里，我突然意识到，可以将米哈伊尔的行为表征理论与双加工理论整合起来。

环线铁路困境确实特殊，这个场景中，伤害确实是达成目标的手段，但其复杂程度却超过了其他的普通事例。具体来讲，在这个场景中，我们必须同时记下多条因果链，才能判断出受害者的伤害来源于达成目标所需的手段。需要掌握的信息已经在图9.8中列出。

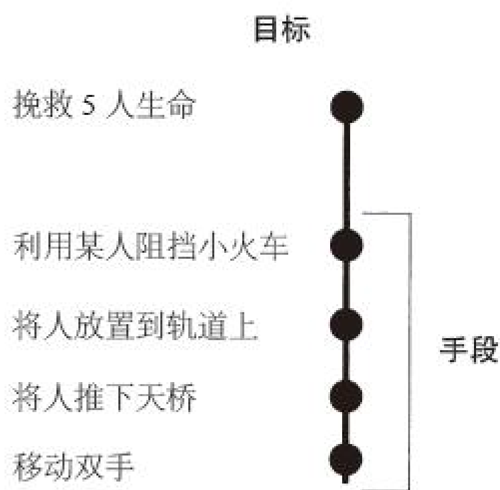



图9.8 人行天桥困境行为计划的主要因果链

也就是说，从肢体动作（移动双手）到最后目标达成（挽救5人生命）的过程中，利用某人阻挡小火车是必不可少的一个有害事件。但在环线铁路困境中，为了得出有害事件必不可少的结论，我们需要同时记录两条因果链。因为小火车在两种情况下都可能撞到5位工人：（1）沿着主线铁路一

直行驶；（2）绕过支线铁路。为了挽救5人的生命，必须同时破坏上述两条因果链。破坏第一条因果链的过程如图9.9所示。

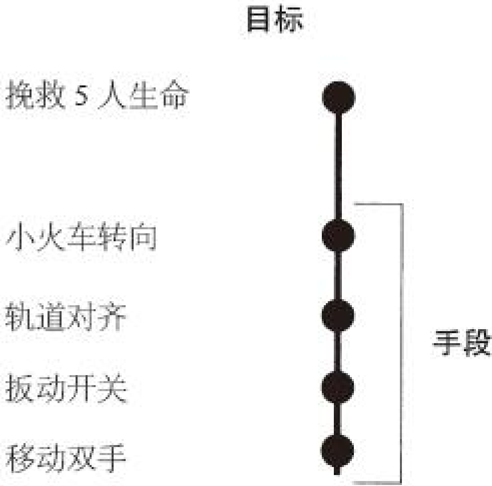



图9.9 环线铁路困境和开关困境中的主要因果链。扳动开关能够避免另外一条因果链的实现，阻止小火车撞上5位工人

环线铁路困境中，小火车向5位工人行驶过去，但通过改变其行驶方向，便能够避免惨剧。也就是说，让小火车转向便能破坏第一条因果链，阻止小火车沿着主线铁路继续行驶，避免撞上5位工人。开关困境中，关于达到目的必需步骤的讨论可以就此结束，因为小火车转向后，无须再做什么，5人便能得救。也就是说，可以将图9.9视为开关困境中救人所需步骤的图示。但在环线铁路困境中，图9.9中标注的事件并不完整。为了挽救5人生命，我们必须破坏第二条因果链。行驶方向改变后，小火车驶上了支线铁路，从另一个方向对5人生命再次造成威胁，形成了新的因果链。为了消除小火车第二次撞上5人的可能性，必须有障碍物在支线铁路上挡住小火车。于是，倒霉的受害者又出现了，破坏第二条因果链的过程如图9.10所示。

如果只看图9.9中标示的主要因果链，便不会发现任何伤害行为。小火车成功转向，没有驶向5位工人，结果很好。要想发现环线铁路困境中的伤害行为，意识到要想达到目标就必须造成伤害，我们需要看到图9.10中标注出的次级因果链。

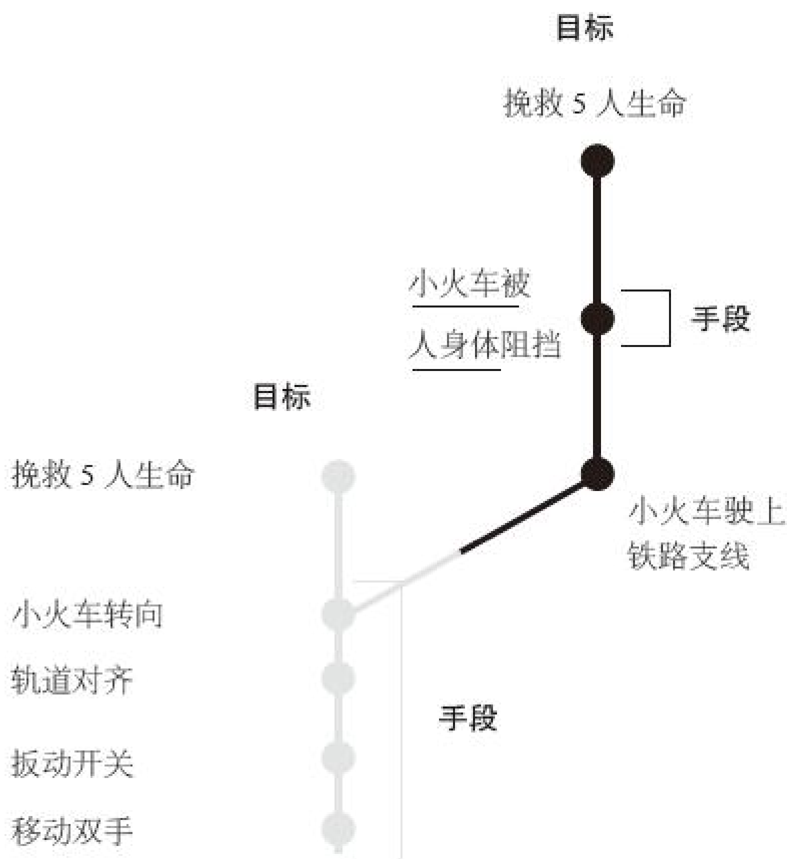



图9.10 环线铁路困境中的次级因果链

让我们先暂时回到双加工理论，该理论认为，大脑中存在一种自动设置，会对人行天桥困境中的行为以及其他类似的有害行为发出预警。而手动模式在自然状态下的思考往往基于成本效益分析。面对开关困境、人行天桥困境以及环线铁路困境等所有场景，手动模式会说“一命换五命吗？听上去很划算”。一命换五命的场景中，手动模式得出的结论永远相同，（“很划算！”）因此这些价值判断最终都由自动模式做出，也就是说，我们做出何种价值判断最终取决于近视模块是否发出警报。*

警报的发出又是由什么决定的呢？我们已经知道，至少在某些情况下，道德判断系统对手段与连带作用的分别十分敏感。但环线铁路困境表明，道德判断系统并不总能感应到这种分别。这是怎么回事呢？此前，我们似乎

找到了合适的答案：因为缺少个人力量的影响。我们之所以反对将人推下天桥，却支持让小火车转向环线，是因为将人推下天桥涉及推的动作。但如果这个解释是全面的，那么在环线铁路困境中加入推的动作，这种功利主义行为是否就会与将人推下天桥同样恶劣？答案似乎是否定的。**既然如此，那么是否还有其他原因，使大脑将环线铁路困境中的伤害形式归入连带作用，而不是按照真实情况，将其纳入因达成目标所用手段而造成的伤害类型？

这里给出一个提示：根据双加工理论，发出情绪预警的系统较为简单，只要人类出现实施暴力行为的想法，行为监控系统便会发出警报。

还有一个提示：如上所述，环线铁路困境是一个异常复杂的手段伤害事例。要想对其正确归类，仅仅关注图9.9所示的主要因果链是不够的，要想意识到伤害是目标达成过程中必不可少的步骤，就必须看到图9.10所示的次级因果链。

你明白了吗？如果我是正确的，那么对环线铁路困境这个谜题，应当做出如下解释：人类拥有一个对行为计划进行“审查”的自动机制，一旦发现伤害事件（比如利用小火车将人撞倒）便会发出警报。但是，（鼓声请响起……）这种行为计划审查机制较为简单，是一种“单通道”机制，无法对多条因果链同时监控。也就是说，这种机制无法关注支线行为计划，当该机制接受行为计划，开始审查时，它只能对发生在主要因果链上的事件保持关注。

这种机制为何采用了这样的工作方式？请想一下自己记忆歌词的方式：《我在铁路工作过》歌词的第三行是什么？就算你能够回答出来，可能也需要一些反应时间。你需要从头开始，一句一句往后想：我在铁路工作过，在我有生的每一天。我在铁路工作过，只为能够消磨时间。你听到汽笛鸣响了吗？……你并不会同时回想整篇歌词，而是从前往后逐句回想，希望每句歌词都会引出下面的一句。我的意思是，大脑潜意识里对行为计划进行监控的机制，与有意识回想歌词时的思维方式相同，都是按照链状结构逐步进行的。对行为计划进行审查时，审查机制从肢体运动开始（比如，推），就像你从“我在铁路工作过”开始一样。随后，审查机制逐项排查，直到最后目标为止（比如，挽救5个人的生命），就像你逐句顺接歌词，直到歌曲结束为止。行为监控机制无法看到行为计划中次级分支上的事件，因为它的工作方式就是沿着主线自下而上逐步排查。

对人行天桥困境的行为计划进行审查时，大脑无法看到图9.7中右半部分的内容，它能看到的只有图9.8所示的内容，但这些内容已经足以触发警报，因为用小火车撞倒工人这项有害行为赫然地列在了主线当中。

然而，对开关困境的行为计划进行审查时，大脑无法看到图9.7中左半部分的内容，它能看到的只有图9.9所示的内容，其中并不包含有害行为。对于大脑的审查机制来说，开关困境中发生的事件只包括：移动双手→扳动开关→轨道对齐→小火车转向→挽救5个人的生命。也就是说，大脑只能看到图9.11中的内容。

幸福的大脑审查机制只能看到主要因果链上的事件，不会意识到通过审查的行为将会导致某人的死亡。由于伤害因连带作用导致，是发生在次级因果链上的事件，因此警报永远不会响起。

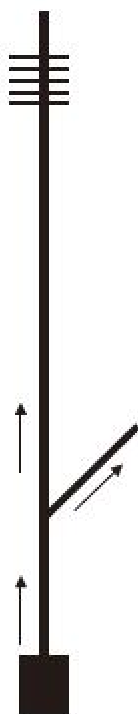


图9.11 开关困境和环线铁路困境中主要因果链所对应的空间事件

那环线铁路困境呢？我们在环线铁路困境中的异常反应恰恰为模块近视假设提供了关键（尽管只是初步的）证据。在这个场景中，之所以说我们的反应“异常”，是因为场景中利用人体来阻挡小火车，伤害来源于达成目标所需的手段，但多数参试者依然支持这种行为。环线铁路困境的一个特点是，它的设计与连带作用伤害事例十分类似，但其实质却是手段伤害事例。具体说来，环线铁路困境中，尽管造成伤害是挽救5个人的生命、达

成最后目标的必需事件，是因果链上必不可少的一环，但有害事件发生在次级因果链上，与连带作用伤害事例相同。前面分析得出，环线铁路困境中的主要因果链与开关困境中的主要因果链完全相同。因此在环线铁路困境中，行为计划审查机制看到的内容只有：移动双手→扳动开关→轨道对齐→小火车转向→挽救5个人的生命。与开关困境一样，因为主要因果链上并不包含有害事件，因此警报永远不会响起。（你可能会问，包含伤害行为的因果链为什么是次级因果链呢？**）由于有害行为不在主要因果链上，所以也无法被近视模块所识别。也就是说，虽然环线铁路困境的实质是手段伤害事例，但其结构与连带作用伤害事例十分相似，所以近视模块机制无法对其识别。*

根据这个理论，近视模块之所以近视，是因为无法识别连带作用，但这并不意味着人类大脑无法识别连带作用。相反，我们完全能判断出，开关困境属于连带作用伤害事例，而人行天桥和环线铁路困境则属于手段伤害事例。如果我们能够看到连带作用，近视模块却无法看到，那就说明，大脑中必然存在另外一个能够识别连带作用的区域。连带作用在大脑中对应的区域是哪里呢？

请再次回想道德判断的双加工理论。近视模块不过是一种自动设置，是决定何时鸣响道德警报的小装置。但双加工理论还包含另外一个方面，即大脑的手动模式。前一章提到，手动模式的设计是为了适应尽量多的用途，为了进行成本效益分析，连带作用自然逃不出它的视线。手动模式能够判断出一件事是否属于连带作用，但它对此并不敏感。诸如用小火车撞倒一个人这样的事件究竟属于手段伤害事件还是连带作用伤害事件，手动模式并不在意（除非我们去攻读哲学系研究生）。手动模式只偏爱最优方案，对于这些事例都会给出同样的答复："一命换五命吗？很划算。"只要警报未被触发，没有相反的观点提出，手动模式就会进行主导，做出判断。这也就是为什么我们会在开关困境和环线铁路困境中给出功利主义答案，而在人行天桥困境中给出相反的答案。

为了便于理解，我们将整个模式画在了图9.12中。

近视模块对行为计划进行审查，一旦发现有害行为，便会发出情感警报。但近视模块无法看到有害的连带作用，因为其审查范围仅限于达成目标所必需的节点事件，也就是行为计划中主要因果链上的事件。近视模块不会针对开关困境和环线铁路困境发出警报，因为它无法看到两个场景中的有害行为（第二行）。但针对人行天桥困境，近视模块会发出有力的警报，因为这个场景中的有害事件恰恰位于主要因果链上（第四行），该有害事件对近视模块是可见的。手动模式能够表征达成目标所必需的节点事件，也能够认出因连带作用而造成的伤害。但手动模式并不"在乎"伤害是因达成目标所需手段造成还是因可预见的连带作用造成的，也就是说，手动模

式不会对因达成目标所需手段造成的伤害产生更多的情绪反应。手动模式关心的底线只有一个：何种行为的最终结果更好？因此，手动模式对一命换五命的行为都会感到满意（第一行和第三行）。近视模块与手动模式以如下的方式进行互动：如果情绪警报没有响起，手动模式就自行其是（第一行和第二行）；如果情绪警报被触发，手动模式的逻辑就会失效（第三行和第四行）。（请注意，手动模式的逻辑不一定总会失效，如果成本效益分析有足够的说服力，手动模式也可能会取代情绪反应，坚持依照成本效益分析的结果做出决定。）因此，近视模块理论将米哈伊尔的行为计划理论与道德判断的双加工理论结合起来，解释了我们对连带作用产生的伤害不太关心的原因。

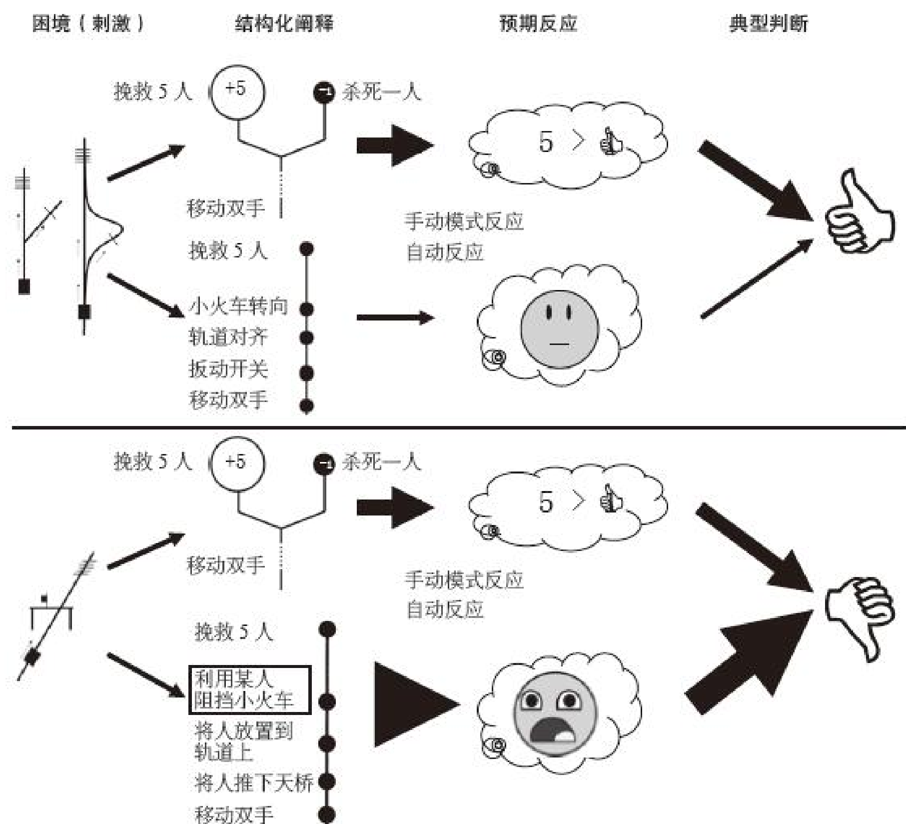



图9.12 双加工机制对开关困境、环线铁路困境以及人行天桥困境的反应。三个困境都触发了手动模式的功利主义思考（第一行和第三行），但只有人行天桥困境中的有害行为触发了针对某些有害行为的自动负面情绪反应（第二行和第四行）

近视模块理论基于一个假设，假设近视模块只能看到达成目标所必需的节点事件，也就是行为计划中主要因果链上的事件。但近视模块为何采用这种工作原理？如果它能同时看到次级因果链上的事件，不是更好吗？若果真如此，近视模块的功能也许会更加完善，但人类的认知能力也必须相应提高。原因有二，其一，迄今为止，所有的讨论都只包含主要的和次级的两条因果链，这个模型已经经过了极度简化，如果不是异乎寻常、极度复杂的情况，在动作执行者的肢体动作和最终目标之间只可能存在一条因果链，也就是说，主要因果链只有一条。但对于一项给定的行动，从肢体动作开始，可能会有很多向外辐射的次级因果链。例如在开关困境中，扳动开关会导致支线铁路上的1位工人死去。但我们还能预见到随之发生的很多其他事件：支线铁路周围的空气受到扰动；受害者家人和朋友的生活会发生变化；扳动开关的人记忆中会留下这些事件；扳动开关的人可能还要面临法律方面的问题，等等。任何行动都会产生无数个可预见的连带作用，如果一个模块想要在关注某个特定结果的同时，还对其连带作用保持关注，它的工作负荷便会成倍增加，因而影响工作效率。但该模块的功能决定它的工作必须是高效的。

其二，对所有支线进行行为计划审查并非易事。要想达到这一目的，我们需要更加复杂的记忆体系，类似于计算机科学中所说的“队列”，即按照顺序将待处理项目存储起来的系统。我们已经介绍过，近视模块对行为计划的审查是线性的，同我们回想歌词的方式一样，链条上的每个环节都能引出下面的环节。因此，近视模块只能对真正的线性链条进行审查。如果链条出现分支，我们就需要存储空间，将一项任务暂停（审查第二条分支），并记得在另一项任务（审查第一条分支）完成后回到这个节点。这种嵌套式的多任务处理，包括对下级任务和上级任务的处理，对于电脑来说十分轻松，但对动物的大脑来说却是极大的挑战。人类大脑的手动模式能够胜任这样的工作，但对于自动模式下的简单认知模块来说，这样的任务异常困难，甚至是不可能完成的。

因此，假设近视模块无法看到连带作用并非毫无依据，相反，这个假设完全符合认知规律。高效自动的行为计划审查机制很难一一“审查”所有可预见的连带作用，因为一个行为引起的连带作用实在太多，完成这项任务所需的记忆系统也实在过于复杂。

我们已经在理论层面进行了很多讨论，最终的结论是：如果模块近视假设是正确的，那么在因手段而导致的伤害和连带作用产生的伤害之间，我们所做出的直觉道德区分也许不过是一个认知意外，一个副产品而已。因手段而导致的伤害会触动道德情感神经，并非因为这种行为在客观上是错误的，而是因为阻止人类滥用暴力的警报系统在认知层面无法对连带作用进行监控。后文还会对此进行讨论，但在此之前，让我们先来探讨另一个反

对最大化幸福感的经典道德区分。

## 执行与允许

上大学时，我有一次在走出食堂时想要扔掉餐巾纸。我把餐巾纸揉成一团，用力抛向已经满溢的垃圾箱，餐巾纸被弹了出来，落在了地上。作为一个有素质的人，我并不想让垃圾箱变得更脏更乱，于是我走上前去，想要捡起餐巾纸，把它放入垃圾箱。但地上丢满了皱成一团的餐巾纸，我也不知道究竟哪块才是我的。我尴尬万分地盯着散了一地的餐巾纸，想根据我那块餐巾纸的飞行轨迹，找到它的着陆地点。过了好久，我终于意识到了自己的愚蠢。我想，一块餐巾纸不过是一块餐巾纸而已。扔在地上的餐巾纸中哪块是我的，哪块是别人的，这完全不重要。我随便捡起一块餐巾纸，将它放进了垃圾箱。但我又想到了一个新问题：为什么只捡一块呢？我已经跨过界线，从“收拾自己的垃圾”走到了“收拾别人的垃圾”，既然我已经捡了一块别人的黏糊糊的餐巾纸，那为什么不再多捡几块呢？可究竟应该捡多少呢？我捡了一把餐巾纸，把它们放进垃圾箱，然后离开了。

我们坚信，相较于以其他方式出现在垃圾箱周围的餐巾纸，我们对自己扔的餐巾纸负有更大的道德责任。这种观点的形成经历了一段独特的哲学历史，被称作“执行与允许原则”。其内容是，因行为和我们实际所做事件所造成的伤害，比因疏忽而造成的伤害更加恶劣。在直觉层面上，这个观点十分具有说服力，在真实世界的道德决策过程中也占据重要地位。例如，美国医学会的伦理规范规定，一名医生永远不能主动、有意地造成病人死亡，但在某些情况下，医生可以有意允许病人死亡。对这种区别的感知也影响到了我们在能够避免痛苦的条件下对痛苦本身的态度。你不会主动引起地震，夺走人们的生命，但你可能会消极对待救援工作，导致地震中的受害者死去。你可能不会在卢旺达或达尔富尔杀人，但你可能并未对他们的防御给予支持，导致他们命丧他人之手等。

从功利主义者的角度来看，执行和允许的区别与道德无关，至少不能将这种区别作为独立的道德力量。我们认为，伤害就是伤害，主动造成的伤害与放任不作为产生的伤害在道德上没有本质区别。（但在实践中，非本质的差别依然存在。稍后会进一步解释。）根据道德观和我们所处的情况，在实际执行的事件和我们任其发生的事件之间画一条区分界线是否有意义？和手段与连带作用的区分一样，我相信我们能基于基础的认知机制对这一问题进行解释，而这种认知机制并不关乎道德。也就是说，我认为对执行与允许之间的区别进行阐释的道德权威是能够被推翻的。

让我们暂时忘却道德。动物的大脑为什么要将主动行为导致的事件和不作为的放任下自然发生的事件区分开呢？阅读这本书时，你主动让双眼扫过

书页，主动翻动书页，等等。这是你实际所做的事情。但考虑一下你没有做的事：你没有教狮子狗跳舞，没有给罗德·斯图尔特写有趣的信，没有摆弄燃烧的火把，也没有在地下室安装热水浴缸。但这只是个开始。在每个时刻，你没做的事情都有无数件，大脑不可能将它们一一列出，即使列出一多半也是不可能完成的任务。（听上去很熟悉吧。）这就意味着在某种意义上，大脑必须重视行为，而不是重视所忽视的事情。大脑需要对某个事件进行表征，随后才能执行，才能确保该事件的进行符合预期计划，才能理解他人的行为。但我们无法对自身和他人没有做的所有事件保持关注。这并不意味着我们无法对被忽视的事件进行考虑，只不过是大脑表征行为事件和表征疏忽事件的方式有本质的不同，对于行为事件的表征更加基础，更加容易。

从婴儿身上可以看出，大脑对行为事件的表征更加基础。菲耶里·库什曼（Fiery Cushman）、罗曼·费曼（roman Feiman）和苏珊·凯里（susan Carey）设计进行了一项开创性试验，对区分行为与疏忽的认知根源进行探索。他们组织6个月大的婴儿观察试验人员对成对物品的挑选，希望婴儿能够发现试验人员的喜好。比如，婴儿的右边放有一个蓝色杯子，左边放有一个红色杯子，试验人员选择了蓝色杯子。下一次，婴儿右边依然放有一个蓝色杯子，左边换成了绿色杯子，试验人员依然选择右边的蓝色杯子。试验反复进行多次，左边的杯子每次颜色都会不同，试验人员每次都选择右边的蓝色杯子。在之后关键的测试选择中，蓝色杯子被放到了左边，右边杯子的颜色与之前任何杯子的颜色都不相同。在一些婴儿面前，试验人员会选择左边的蓝色杯子，而在另一些婴儿面前，试验人员会选择右边的杯子。关键问题是：哪一种选择更加出乎婴儿的意料？一方面，试验人员一直都在选择蓝色的杯子；但另一方面，试验人员一直都在选择右边的杯子。婴儿会因此对试验人员产生怎样的预期？选蓝色还是选右边？试验表明，试验人员选择右边的非蓝色杯子时，受到婴儿注视的时间更长，表明婴儿对此种行为感到惊讶，也就是说，根据婴儿的预期，试验人员会选择蓝色的杯子。这个试验表明，6个月大的婴儿大脑中表征的事实是：试验人员想要蓝色的杯子。婴儿们做出的动作也与大脑表征相一致。

这只是试验的一半，另一半的试验中，研究者组织婴儿们多次观察试验人员在蓝色杯子和其他颜色的杯子之间进行选择。不同的是，这一次试验人员总会选择其他颜色的杯子，而不是右边的蓝色杯子。在随后关键的测试选择中，试验人员依然要在左边的蓝色杯子和右边新出现颜色的杯子之间选择，我们的问题依然与婴儿的预期有关。第一次试验中，婴儿看到试验人员多次选择蓝色杯子后，便会认为试验人员还会选择蓝色的杯子。但这一次，婴儿看到的是试验人员多次没有选择蓝色杯子（多次的忽略），他们会认为试验人员将不选蓝色的杯子吗？

不会。事实上，试验人员选择蓝色杯子时，婴儿们没有表现出任何惊讶的迹象，也就是说，婴儿们能够掌握“选择蓝色杯子”的概念，却无法掌握“不选蓝色杯子”的概念。请注意，婴儿有能力区分对蓝色杯子的“选择”和“不选择”，因为如果他们无法区分这一点，那么起初的试验中，试验人员多次选择蓝杯子后突然选择非蓝色的杯子时，他们便不会惊讶。这只能说明婴儿建立了某种预期，并且发现实际情况与预期不相符。很明显，婴儿们的大脑不会表征“不选蓝色杯子”，也不会将其作为特殊的行为。对试验人员进行观察后，他们自己不会想到“看，他又一次没选蓝色的杯子”。

这项试验表明，在大脑中表征“选择蓝色的杯子”等以目标驱动的行为属于基本的认知能力，6个月大的婴儿便已经具备了这样的能力。但是，对疏忽行为或未做某个特定事件的表征则是相对高级、复杂的能力。需要注意的是，表征疏忽行为并不一定需要极其复杂的信息处理能力。如果总共只有两种可能性：选a或是不选a，那么对未完成事件进行表征的难度和对已完成事件进行表征的难度便相差无几。如果我们用电脑编程、监控并预测某人对杯子的偏好，假设某人的选择只有两种，那么在程序中要求电脑显示某人“未选蓝色杯子”和“已选蓝色杯子”的难易程度几乎相同。（只需一个“非”算符，就能将后者变为前者。）尽管如此，人类似乎依然认为表征已完成事件比表征未完成事件容易。这种偏好是有道理的，现实生活中，人们能做但没做的事情成千上万，相较之下，对为数不多的已完成事件保持关注才更加重要。

婴儿对已完成事件的表征比对未完成事件的表征更加容易，这个事实可以帮助我们预测成人的行为：成年人在道德判断中区分有害的行为和有害的疏忽（未完成事件）时，将执行与允许原则付诸实践的是自动模式而非手动模式。库什曼和我在一项脑扫描研究中共同验证了这项预测。研究当中，我们要求人们对积极有害行为和消极有害行为进行评价。结果发现，在道德上将两种有害行为等同起来，也就是将行为与疏忽的区别忽略时，负责手动模式思维的背外侧前额叶皮层会更加活跃。**这种现象是合理的，因为对疏忽的表征本身就是更加抽象的思维。与疏忽不同，行为本身能够通过基本的感觉方式进行表征。比如，画一幅某人奔跑的画十分容易，但要想表达某人没在奔跑，该如何落笔呢？你可以画一个静止站立的人，但人们会认为你画的是“人”、“女人”或“站立”等其他概念，而不会想到你画的是“没在奔跑”。表征“某事没有发生”的传统方式是使用抽象符号，比如一个圆圈，里面画一条斜线，再配以一个传统的图像。但仅靠传统图像无法表达这一概念，抽象符号是必不可少的。

行为不仅能通过自然的感觉方式进行表征，还能通过自然的运动方式进行表征。朗读“舔”、“捡”或“踢”等词语时，运动皮质中控制舌头、手指和脚的亚区活动便会自动兴奋起来。但当人们思考“不包括舌头”的行为时，大

脑中所有区域的活动都不会增加，因为大脑中没有哪个区域专门负责“不包括舌头”的行为。

如前所述，人类的感情以及道德判断，似乎都对行为的感觉和运动特性更加敏感，比如推的动作等。（对推的画面也同样敏感。）与行为不同，大脑中没有与疏忽相对应的感觉和运动特征，这方面欠缺的情感触发机制也许不止一种。此外，行为与疏忽在感觉和运动方面的区别可能会影响到更加无序的肢体行为，我们的认知程度也与此有关。例如，“解雇”某人（主动）比“允许某人离开”（被动）感觉更糟糕。与这个结论相呼应，尼尔鲁·帕哈里亚（Neeru Paharia）、卡里姆·卡萨姆（Karim Kassam）、马科斯·巴泽曼（Max bazerman）和我曾经做过一项研究，结果表明，通过第三方代理的方式间接提高癌症药物的价格会减轻人们心理的内疚感，尽管这种行为本身一点都不间接。

因此我们的假设是，有害疏忽对人类情感道德神经的触动并不像有害行为那样强烈。我们能以基本的感觉方式和运动方式表征行为，但对疏忽的表征则更加抽象。此外，人类表征行为和疏忽的不同方式与道德无关，不同的表征方式源于大脑本身的认知局限。最初大脑作为感觉机制和运动机制，而非抽象思维机制进行进化，无法对未进行的所有行为进行关注。我们再次发现，一直被奉为经典的道德区分也许不过是大脑认知的副产品。（稍后会解释，功利主义依然能对行为与疏忽的区分进行调节。）

## 功利主义与机制

将幸福感最大化听上去是个绝妙的想法，但至少在理论层面上，它可能意味着做出极其糟糕的事情。该怎么办呢？道德思维告诉我们，“理论上的”问题可能根本不是大问题。“理论上的”问题并非不值得担心，而是因为有证据表明，直觉对值得警惕的事件往往会做出可靠的反应。由于大脑采用了道德思维双加工机制，很多好事都会让人产生大错特错的感觉；也会有很多坏事给人的感觉似乎没什么大不了。为了对警示行为进行总结阐释，还是要回到我们最钟爱的果蝇——人行天桥困境上来，然后再探究讨论结果的深远意义。

尽管将人推下天桥能够挽救更多生命，但这种做法似乎依然不对。为什么呢？我们从第4章得知，这是大脑自动模式的工作结果，但这并不是答案的全部。基于认知机制的工作特点，我们可以得到更加完整的答案：这项机制会对什么做出反应？对什么不做出反应？让我们先从第一个问题开始。

首先，自动模式对因达到目标所需手段造成的伤害更加敏感，对因连带作用导致的伤害则没有明显反应。（如果有害行为看上去类似因连带反应造

成的伤害，则反应也不明显，这是该模式奇异的特点之一。）也就是说，自动模式对伤害的反应有针对性。*其次，自动模式对主动造成的伤害更加敏感，对于被动造成的伤害，反应则没那么明显。最后，对于通过个人力量导致的伤害，反应更加明显；对于间接的伤害，反应则没那么明显。这三个方面似乎不是像清单一样的、相互独立的标准，相反，在警报机制运转过程中，三个因素似乎相互交织，构成了有机的整体。个人力量和手段与连带作用的区别互相影响：如果伤害不具有针对性，那么是否由个人力量造成便无关紧要；如果伤害由个人力量造成，那么这种伤害是否具有针对性，意义便不太大。此外，主动与被动造成的伤害与其他两个因素也相互交织。人们对自己在小火车问题中做出的直觉判断进行解释时所犯的错误便恰好可以说明这一点。

为什么人行天桥困境中将人推下天桥是错误的，但在开关困境中扳动开关的行为却是正确的？人们回答这个问题时，通常会用行为与疏忽的区别进行解释，但事实上这个理由并不成立。人们会说："将人推下天桥是谋杀，你杀死了一个人。但在另一个场景中，你只是看着他被小火车撞死。"这个解释是说不通的，两个场景中，伤害都是主动发出的。假设你抱着杀死某人的目的，让小火车向他开去，这种行为绝对是主动的，绝对是谋杀。开关困境中，我们做出的动作与谋杀并无两样，也并没有更加被动，但因为开关困境中的伤害以连带作用的形式出现，并且没有个人力量的直接参与，给人的感觉像是被动。因此，三种因素似乎同时都在对同一种感觉施加影响。

这种情况和描述手段与连带反应区别的认知机制理论刚好相符。根据模块近视假设，有害的连带反应不能触发警报，因为有害事件并没有位于行为计划的主线之上。但被动行为造成的伤害并没有相应的行为计划（通常情况下），因为动作并非主动发出，因此不会有肢体运动，也没有连接肢体运动和最终目标的事件链条。行为计划理论能够解释我们为何对手段与连带作用的区别如此敏感，也刚好可以解释我们为何对行为与疏忽的区别如此敏感。*

个人力量也会对行为计划产生影响。行为计划中的事件并非随机排列，而是按照因果关系进行排列。从肢体运动开始直到最终目标达成（扳动开关……小火车转向……挽救5人生命），每个事件都是下一事件的起因。有证据表明，人们习惯用力来表征起因。当你看到两个台球相互撞击时，视网膜上呈现出的图像显示了台球的一系列位置，就像电影里的一系列画面一样。但不管怎样，我们凭直觉确切无疑地知道，力从一个台球转移到了另一个台球上。因此，行为计划中表征出来的力，不论是个人力量还是其他形式的力量，可能会影响到思维中个人力量的施加与造成伤害之间的关系。当然，个人力量的应用与行为和疏忽之间的区别也有关系，因为疏

忽本身并不包含个人力量。

将三个特征放在一起就会发现，我们的警报机制会对典型的暴力事件做出反应，如打、掌掴、拳击、棒打，当然还有推。*还有一些其他行为，并不具有上述三个特征，但也会造成伤害，例如不为慈善事业捐钱，因而不能救人性命等。但这些行为却一点都不暴力。同样的，如果某种行为通过个人力量的介入，以造成伤害为目的，作为达成目标的必需手段，主动造成了伤害，那么这种行为几乎不可能不包含暴力。*我并不认为大脑的自动警报系统由暴力行为而产生，相反，我怀疑这种自动警报系统是人类定义暴力的最初依据。

我们已经就警报机制的触发事件进行了讨论，但会被警报机制所忽视的是哪类事件呢？所有事件当中，警报机制会对暴力行为带来的好处视而不见。我的搭档和我假设将人推下天桥的行为能够拯救上百万人口，提出了另一种人行天桥场景：如果小火车继续行驶，它将会在经过大坝顶端时撞上一箱炸药。如果堤坝被炸毁，整个大城市将被洪水淹没，上百万的人将会因此丧生。这个场景中，70%的参试者支持将人推下天桥。尽管这个场景中将人推下天桥所获得的好处比开关困境增加了几百万倍，但其支持率依然低于开关困境中87%的支持率。因此，警报机制似乎并不“在乎”处于危险境地的其他事物。当然，与只有5人生命受到威胁的原始人行天桥困境相比，对将人推下天桥这种行为的支持比例确实提高了很多。显然，人类的判断会受到数量的影响，但这样的判断结果似乎并非因为警报没有响起，而是因为面临巨大的数字，大脑选择忽略情感警报，或是推翻了警报的结果。这个道理细想便很清楚：将人推下天桥来挽救一百万人和挽救5个人相比，并不会让人感觉更好。相关的试验也证明了这一点。不愿相信直觉（进行更多“认知型反思”）的人们当中，更多人选择支持将人推下天桥，以挽救百万生命。试验结果表明，尽管大多数人都对这种行为表示支持，但这个选择确实与直觉相悖。

至此，我们已经对导致警报机制作为和不作为的事件进行了充分了解。简而言之，警报机制会对典型的暴力行为做出负面回应，而忽视这些行为可能带来的积极后果。基于这项结论，究竟应以何种态度对待警报机制向我们传达的信息呢？

总体来说，我认为我们应当严肃对待警报机制的提醒。暴力行为通常是不好的，因此每当我们想通过暴力手段达到目标时，大脑中的警报机制便会向我们尖声报警。这是一件好事，如果没有这个机制，人类也许会更加变态。此外，警报机制还能预防过于自信的态度，避免偏见的产生。即使你考虑使用暴力时心中怀着最好的目的（“这场革命也许十分血腥，但想想我们光辉的未来吧！”），警报也会尖声叫道：“小心！你在玩火！”大脑中有这样一个声音是一件好事。总之，一想到要将无辜的人推向死亡，警报

系统便会让我们心生畏惧，大体上讲，这是非常好的。

但不论大脑中的反暴力机制多么必不可少，我们都不必将其结论视为绝对真理，也不必将其工作机制上升到道德原则的高度。警报机制能够辨别手段与可预见连带作用的区别，并非因为这种区别本身反映了道德价值，而是因为警报机制的认知能力有限，无法对多条因果链进行关注。同样，警报机制也能区分主动伤害与被动伤害，这并非因为主动伤害的本质比被动伤害恶劣，而是因为该机制的设计目的就是评判行动计划，而大脑对行动与非行动的表征方式并不相同。最后，警报机制也许会对有个人力量介入的伤害做出更大的反应，这并非因为个人力量本身十分重要，而是因为人们可能会做出的基本邪恶之事（打、推等）大多都会使用个人力量。

这些区别并非与道德全然无关。一方面，由于大多数人都会意识到这些区别，我们可以据此研究不在意这些区别的人，对他们的道德特征进行推导。如前所述，有重要证据表明，对愿意通过个人力量造成伤害的人们来说，他们的反暴力警报机制可能存在某些缺陷。道德观念正常的人不会做出这样的行为。如果某人缺乏正常的道德观念，那么他的道德观念很可能是有缺陷的。（道德观念完好但不正常的情况极其少见。）对想要故意导致伤害和主动造成伤害的人来说，情况也是如此。然而，所有这一切都能被功利主义思想所融合：对人进行评判时，认真考虑行为与疏忽的分别、手段与连带作用的分别以及是否使用个人力量的分别是合理的行为。并非因为这些分别能够反映深层道德真理，而是因为忽略这些分别的人们道德意识不正常，更可能为自己惹祸上身。

行为与疏忽的分别对功利主义思想十分重要。如果没有这项分别，你就需要对自己有能力避免的所有麻烦负责，好像你是所有麻烦的制造者一样（餐巾纸，到处都是餐巾纸……）我们不是超人，无法以解决世界上所有问题为己任，所以合理的状况是，人们只对自己的行为承担特别责任（把你自己的餐巾清理干净！）。同时，请回想我们将行为与疏忽的分别归类为直觉思维的原因，两种解释都基于同一个事实，即疏忽事件的数量远远多于行为事件。即使是一瞬间，大脑也不可能记下这一瞬间没有发生的所有事（疏忽）。同样，我们也不可能为未进行事件所导致的麻烦负责任。当然，主动造成伤害与被动造成伤害在本质上并没有优劣之分。（一块餐巾纸不过是一块餐巾纸而已！）

因达到目的所需手段造成的伤害与因连带作用造成的伤害之间的直觉区分对实际生活具有重大意义。将有针对性的伤害与可预见的连带作用造成的伤害区分开来并没有太大的必要，但将有针对性的伤害与因不可预见的连带作用造成的伤害（意外）区分开来却十分重要。无意中伤害到他人的人也许是危险的，但将伤害他人作为手段达成目标的人才是真正的危险。与明知行为的后果，还依然对他人造成附带损伤的人相比，这些人也许更加

危险，也许不相上下。但针对有意伤害发出道德警报至少可以将“不择手段”的伤害与意外伤害区分开来，即使警报系统没能正确归类，将可预见的连带反应也视作意外，警报的存在也是一件好事。

因此我依然认为，我们应当为人类大脑中反暴力机制的存在而感到庆幸。但关键问题在于：我们的首要道德哲学是否应由反暴力机制决定？我们是否应当听从该机制的指挥，放弃追求更大范围的利益？我们是否应当从人行天桥困境等事例中得出结论：有时候将幸福感最大化是大错特错的选择？或者我们应当得出的结论是，人行天桥困境是一个古怪的例外，并不必为此挂心在意？

我之所以将人行天桥困境称为道德的果蝇，这个比喻有着双重含义。如果我的观点正确，人行天桥困境是道德的害虫。这是一个完全人为设置的场景，这个场景中，典型的暴力事件必然（场景规定）能够导致更大范围的利益。哲学家们从该场景中得出的结论往往是：对更大范围利益的追求有时是大错特错的。然而，依照我们对道德思维双加工理论的理解，可以得出一个不同的结论：人类的道德直觉大体是通晓事理的，但它并非绝对可靠。因此我们必然能设计出各种场景，揭示道德直觉死板的认知能力，也必然能够设计出一种正确的行为，只因为该行为触动了人们的道德神经，便被视为大错特错。我能想出的最好例子无外乎人行天桥困境。

与多数人一样，你也许会怀疑，我是否想证明将人推下天桥是正确的行为。但其实我想说的是：如果你感觉将人推下天桥完全没有问题，那么你肯定是有问题的。我也感觉这种做法不对，也怀疑自己如果真的身临其境，是否真能做出“推”的选择。我对自己的想法感到十分满意。事实上，在现实世界中，不推无疑是正确的选择。但如果有人怀着最好的意愿，鼓起勇气将人推下天桥，确定这样做能够挽救5条生命，并且确定不存在更好的解决方案，那么我会支持这种做法，但我很可能会对这个人心存疑虑。

下一个问题：如果人行天桥困境等场景是古怪的、人为构建的、可以忽略的，那么我们为何要花费如此多的时间对其进行研究？答案是：这些场景有时可以忽略，但有时却不能忽略。要想寻找恰当的元道德，那么就应当忽略人行天桥困境，因为我们不能让因场景而触发的道德警报阻碍我们追求更大范围的利益。但要想了解道德心理学，我们就必须认真研究此类困境。我希望大家能够明白，这些古怪的困境是研究道德思维工作机制的最佳工具。事实上，道德困境的科学意义可以与视觉错觉对视觉科学的贡献相媲美。（视觉科学家对视觉错觉十分推崇，视觉科学学会每年都会评选年度最佳新错觉。）例如，我们十分熟悉的缪勒·莱耶错觉（Muller’s illusion）显示，视觉系统将汇聚的线条作为长度的判断依据。

在图9.13中，位于上方的横线显得比下面的横线长，但事实上两条线的长度相等。视觉错觉能够揭示视觉认知的结构，古怪的道德困境也能揭示道德认知的结构。这些道德错觉通过对我们的误导来揭示真相。

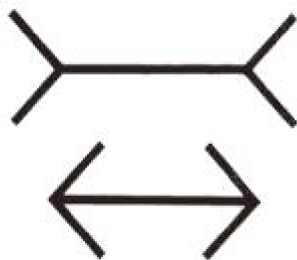


图9.13 在著名的缪勒·莱耶错觉中，两条长度相等的线看上去有长有短

小火车困境也许十分古怪，但它们却真实地反映了现实生活中事关生死的道德问题。人行天桥困境之所以古怪，是因为在这个场景中，典型的暴力行为能够造成更大范围的利益。日常生活中，这种状况很少发生，但在生物伦理学领域，现代知识和科技使人们有机会通过伪暴力行为促进更大范围的利益，这种情况并不罕见。

请回想美国医学会对医生协助自杀采取的立场。从本质上讲，美国医学会对双重效应原则和执行与允许原则持支持态度。如果我是对的，美国医学会所支持的其实是近视模块的工作机制。慢性病人之所以要在病中煎熬，是因为我们没有勇气去主动、有意、人性地做出对他们最好的选择，也是他们想为自己做出的选择。（当然，帮助病人结束生命需要极其谨慎，功利主义完全支持这种谨慎。但美国医学会将医生协助自杀变成了不可逾越的“规范”。）近视模块的工作特点可能还会影响人们对强制疫苗接种、器官捐赠政策及堕胎政策的态度。事实上，小火车问题正是在堕胎问题和双重效应原则问题的讨论中产生的。

在医学领域以外，近视模块的工作机制还会影响我们对死刑、刑讯逼供及战争的态度。所有这些例子中，暴力行为和类暴力行为都触动了我们的道德神经，但它们也都符合更大范围的利益。我们可能会拒绝暴力行为，并不是因为我们已经仔细思考了所有相关的道德考量，而是因为我们的感觉如此。道德警报可能会反应过度，对某些值得警惕的事件带来的好处视而不见；也可能会过于迟钝，忽视某些不值得警惕事件的负面影响。例如，我们因污染环境对他人（包括未来的人）造成伤害时，伤害的方式几乎总是连带作用、是被动的，也永远不会有个人力量的直接使用。如果破坏环境给人的感觉就像将人推下天桥，那么地球可能会比现在美丽得多。*

需要注意的是，从政治角度来看，对警报机制判断能力的怀疑各有利弊。警报机制也许对医生协助自杀和堕胎持反对态度，但它对刑讯逼供和死刑也同样持否定态度。我在这里并未表态支持或反对某项政策，我只想说，我们可以训练自己从不同的角度看待这些问题。

人类对暴力行为保持警惕是一件好事，但大脑中的自动情感机制并不能达到无限智能。道德警报系统认为，将人推下天桥和扳动开关两种行为在道德上有着天壤之别。更重要的是，道德警报系统认为以满足私利为目标的谋杀和牺牲一人挽救一千人之间并没有本质差别。因此在寻找通用道德哲学的过程中，这些机制的判断并不能作为依据，否定我们的选择。

# 第10章 公正与公平

上一章关注的重点是使用值得警惕的手段追求好结果的行为，这一章将重点讨论结果：我们的最终目标是什么？有人说，追求更大范围的利益必然无法兼顾公正，以此作为目标会使我们对他人甚至自己做出不公平的行为。下面我们便就这个观点及其背后的心理学机制进行讨论。

如前所述，我们的策略依然有两个：有时进行协调，阐明在现实世界中将幸福感最大化的做法并不会像某些人所想象的那样，产生荒谬不堪的结论。有时则会提出改革意见，利用我们从认知角度和进化角度对道德心理学的理解，对直觉的正义感提出质疑。

## 功利主义提出的要求过分吗？

如第8章所述，将最大化幸福感的观点身体力行并不容易，因为世界上充满了本可以避免的不快乐。如果想要挽救某人的生命，捐赠2500美元也许可以实现。你可以每年拿出500美元，连续坚持5年，也可以再找4位朋友，每人拿出500美元。除此之外，你还可以通过捐款帮助很多人摆脱悲惨的境地，所需金额并不惊人，也许还不及在餐馆吃一顿饭的花费。简而言之，以恰当的方式花在别人身上的1美元，可能会带来很大的幸福，这种幸福感远远超过你为自己或是家人、朋友花费1美元所带来的幸福。

也许你对这种说法的真实性持怀疑态度。坦率地说，你心里的某个角落也许希望这种说法不是真的。如果你根本无法对世上最不幸的人施以援手，那么便可以放松了。但很遗憾的是，你无法摆脱这种责任，我也一样。确实有一些慈善项目弄巧成拙，造成的伤害超过了所做的好事；也确实有一些寄托着最美好祝愿的善款所托非人，落入了卑鄙独裁者的小金库。但在这个时代，你无法以帮助无门为借口逃避对他人的帮助。国际救援组织比以往任何时候都更加有效，更加可靠。即使有些组织浪得虚名，但只要有一个组织是可靠的，你就无法逃脱责任。可靠的国际救援组织有很多，但即使世界上最好的人道主义组织都将其一半资金肆意挥霍（这种情况当然不会发生），我们依然无法逃脱责任，因为这种情况仅仅意味着提供帮助的成本翻了一倍，并不能从根本上改变我们的计算。当今世界，只要你愿意，就可以为绝望中的人们提供捐款和帮助。

从根本上来说，这种情况是非常好的。但像我们一样拥有一定可支配收入的人却因此而面临艰难的道德选择：100美元就能让一个贫困的孩子连续几月吃饱肚子，你怎么忍心把这笔钱花在非必需的事情上呢？那么下一个100美元呢，再下一个100美元呢？你还能去度假吗？能带某人去约会

吗？你的爱好呢？选择薪酬水平并非最高的事业？为孩子举办生日聚会？甚至从一开始就不应该生孩子？能在比萨上加点配料吗？或者根本不应该吃比萨？对绝对的功利主义者来说，所有这些问题的答案都一样：为了保证你最小限度的幸福，尽可能提高自己让别人，特别是与你无关的人，更加幸福的能力，只要这些享受是必需的，那么便可以安心享受。简而言之，绝对的功利主义者意味着放弃生活中几乎所有的享受，将自己变成幸福之源。

几年之前，一位哲学家在大会上发言，为这种功利主义理想辩护。问答环节时，另一位哲学家站起来，指着发言人的笔记本电脑说："这个电脑至少价值1000美元，而世界上现在还有人在忍受饥荒，你能证明自己行为的合理性吗？"发言人回答道："我不能！但至少我愿意承认自己是个伪君子！"在我看来，这个回答不仅好笑，而且发人深省。（这个答案并不很准确，下文将会解释，发言人完全可以证明自己拥有笔记本电脑是合理的。）

只要我们不要求自己成为绝对的功利主义者，功利主义思想的高要求便不会成为决定性的负面因素。事实上，成为绝对功利主义者的想法十分不符合功利主义思想。回想一下我们想要健康饮食时面临的类似困境：要想摄入绝对健康的饮食，就必须确定最为健康的食物组合，然后便只吃这些食物，并精确控制最佳摄入量。要想保持最健康的饮食，你可能永远不再有机会吃自己最喜爱的食物，即使是过生日也不行。出门旅行时，需要随身携带健康食品，因为你很可能无法在目的地找到这些健康食品。接到朋友发来的晚宴请柬时，你只能拒绝，或者提前吃饱，或者之后再吃，抑或是带着自己的健康食品前去赴宴（天啊！）。你永远无法去饭店与人约会，或者你们约会的餐馆必须提供健康食品。诸如此类的情况数不胜数。

如果你是一台需要进食的计算机，那么保持最健康的饮食是比较现实的目标。但作为一个真正的人，时间和金钱都是有限的，意志力也有强有弱，保持生理学最佳健康饮食这种行为本身便不是最佳选择。相反，最佳的策略应当是在现实世界的限制内，在自己的心理承受范围和社会要求的范围内，尽可能吃好。这样做其实很难，因为不存在神奇的公式指导，在完美主义的极端与放任的暴饮暴食之间也没有清晰的界限。在现实而不仅是理论层面上，要想达到能力范围内最健康的状态，你需要设置清晰的目标，尽管这些目标难免随意，你还是要做出合理的努力争取达到目标。

因此，要想成为现实生活中活生生的功利主义者，道理也是一样的。功利主义理想中的"道德饮食"与大脑机制默认状态下的生活完全不兼容。人类的大脑本身不会过多关注陌生人的幸福，事实上，大脑在默认状态下对陌生人的态度可能是漠不关心，甚至怀有恶意。因此，现实生活中活生生的功利主义者必须放过自己，甚至比现实中提倡健康饮食的人对自己更加宽

容。

在实际行为中，这意味着什么呢？并不存在神奇的公式指导，只是在两个极端之间存在一段模糊不清的恰当区间。生活中的功利主义者不必把自己变成幸福之源，要想究其原因，思考一下后果便不难明白：首先，你根本不会去尝试。其次，如果你真的想要尝试，生活便会变得非常悲惨。早上起床的所有动力几乎都不复存在（如果你还有床的话）。作为一个并不坚定的幸福之源，你可能会迅速说服自己逃离这种哲学，或者索性承认自己是一位伪君子，回到最初的生活状态，重新思考自己究竟多想成为一名伪君子，又有多想成为一名英雄。

然而，做一名生活中的功利主义者并不意味着变成彻底的伪君子，原谅自己的各种行为。无法坚持完美的健康饮食不代表每一顿饭都可以胡吃海塞；同样的，无法成为绝对的功利主义者也不意味着你可以就此免责。很多时候，只要做出相对很少的牺牲，便能帮人减轻很大的痛苦。你该做出多大的牺牲？这里依然不存在神奇的公式，只能根据个人情况和个人局限来决定。我们还可以从社会角度看待这个问题，从长期来看，孜孜不倦的努力比英雄主义行为更加有用。你的生活会成为他人的榜样，特别是你的儿女（如果你有儿女的话）。如果你每年都通过慈善捐助提高几百人的生活水平，与此同时你依然过着幸福舒适的生活，那么你就成了他人可以效仿的学习对象。相反，如果你将自己逼到了崩溃的边缘，尽管你捐赠的金额更多，能够帮助的范围更大，但你的生活毫不令人向往，可能还会对整个慈善事业起到相反的作用。从长远来看，对适度的、可持续的利他主义文化进行鼓励会产生更好的影响，将自己逼迫到极限状态。为他人做出巨大牺牲的英雄是“鼓舞人心的”，但谈到对现实行为的促进，研究表明，让人们去做善事的最好方式就是告诉他们，邻居都在这样做了。

总体的概念是这样的：如果功利主义对你提出的要求看似十分荒唐，那么这绝不是功利主义对你提出的真正要求。如前所述，功利主义是一种本质上十分实际的哲学。要求自由的人们去做他们认为荒唐的事，去做与他们最基本的动机背道而驰的事，这件事本身就是无与伦比的不切实际。因此在现实世界里，功利主义提出的要求确实很高，但并非不切实际。它能够协调基本的人类需求和动机，但绝对没有试图对人类自私的习惯进行大刀阔斧的改革。

当然，有人可能会对现实世界中功利主义思想所提出的适当改革表示反对。有人可能会说，帮助陌生人的行为令人钦佩，但这完全应当出于自愿。这究竟是正当的道德立场，还是温和的合理化建议？为了解决这个问题，我们来研究彼得·辛格（Peter singer）最初提出的道德问题，考虑帮助他人的直觉责任感背后的心理学因素。

## 助人为乐的义务

假设你外出到公园漫步，突然看到一个小孩子掉进了一个浅水塘。你可以轻松涉水过去，挽救孩子的生命，但如果踏入水中，你花费500多美元新买的意大利西装就会毁掉。如果为了保护西装，任由孩子淹死，这是合乎道德的行为吗？显然不是，这样的做法简直是道德上的禽兽之举。但辛格问道，如果你能向国际救援机构捐出500美元，挽救一个孩子的生命，但你却用这500美元为自己买了一套西装，这样的做法合乎道德吗？换句话说，如果你认为挽救落水儿童是道德义务，那么为什么挽救远方的贫困儿童就不是道德上的义务？

（也许500美元不足以挽救一条生命，但衣着考究的同事告诉我，500美元也买不起一身真正时尚华丽的西装。不管怎样，你可以假定有4位朋友，只要你捐款，他们便会跟着捐款；或者可以假定自己会连续4年捐款。）

首先，我们要花一点时间来认可好西装的价值。假设你是一名企业律师，经手的案件价值上亿美元，对你来说，在杰西潘尼百货商店购物就是因小失大的行为。优雅的穿着能够反映你的自信和能力，是安全的投资行为。出于同样的原因，你可以拥有办公室里高档的橡木家具、乡村俱乐部的会员身份、美满的家庭、适宜的娱乐活动，等等。（如果你是一位学者，需要通过阅读和写作谋生，那么拥有笔记本电脑也是合理的行为。）总体说来，不论表象如何，从功利主义的角度来看，很多看似不必要的奢侈品都是必要的。这个观点非常合理，对于包括我自己在内不愿过多改变生活方式的人来说，这个观点很合我们的心意。但辛格的问题依然存在，因为无可否认的是，生活在富裕世界中的我们必然拥有一些可支配的收入。为了保护经济上有用的西装而任由孩子溺水而亡，这依然是道德上的禽兽之举。为什么呢？因为如果你买得起这样的西装，那么你就必然有能力再买一身。如果你救起落水儿童后能够再买一身新的西装，那么购买新西装之前，你也一定有能力挽救一位远方儿童的生命。广义来说，一旦自身的需要得到满足，我们便必须面对道德上的种种可能。

也许我们可以忽略远方儿童的窘境，因为他们是其他国家的公民（至少在我们的故事中是这样的）。但当我们出国旅行时，眼睁睁地看一个外国小孩儿在外国的水塘中溺水而亡合乎道德吗？如果很多人都能为远方的孩子提供帮助，也许我们对孩子的道德义务便得以减轻，而在辛格的落水儿童事例中，你是唯一能够施以援手的人。但这个因素有多重要呢？假设水塘周围还有其他人，他们看到了落水的儿童，但没有试图施以援手，那么我们就能够任由孩子淹死吗？结论是：如果对身旁的落水儿童和远方饥饿的儿童区别对待，想要为此行为找到合理依据似乎极其困难。

不管怎样，显而易见的是，我们必须挽救身旁的落水儿童；同样显而易见的是，向国外援助捐款并不是道德上必尽的义务。换句话说，直觉认为这两件事并不相同。我们应当信任直觉吗？将两件事区别对待是经过深思熟虑后得出的结论吗？或者这只能说明自动模式的死板僵化？

为了回答这个问题，杰·穆森（Jay Musen）和我进行了一系列试验，试图找出类似情况下影响人们判断的因素。我们的试验像是彼得·辛格版的“小火车问题”，*但我们并没有对所有可能的因素进行关注。比如，如果你自己的孩子或者你的侄女也在考虑范围内，你的判断当然会因此不同，但这个因素对解决辛格的原始问题毫无帮助。*我们想知道为什么挽救落水儿童是“必须做！”的事，而抗击世界贫困则是值得钦佩但仍有回旋余地的事。

试验结果表明，到目前为止，空间距离是最为重要的影响因素。我们给出了这样一个场景：你正在刚刚遭受台风猛烈袭击的一个发展中国家度假，幸运的是，你没有受到风暴的影响。你在一座小山上找到了一间舒适的小屋，出门便能看到海岸，生活用品也一应俱全。但在这个国家，救援工作已经展开，你可以进行捐款，提供帮助。在另外一个场景中，身处受灾国家的人变成了你的朋友，其他一切条件都完全相同。而你正坐在家中的电脑面前，你的朋友详细描述了灾区的情况，还用智能手机的照相机和话筒功能，让你从视觉和听觉上对受灾地区的情况进行直观感受，创造出身临其境的感觉，你可以通过在线捐款的方式提供帮助。

在身处灾区的场景中，68%的参试者表示提供帮助是道德义务。相比之下，如果身处距离很远的地方，只有34%的参试者这样认为。在空间距离很远的场景中，尽管其他所有条件都完全相同，提供帮助的能力条件也完全相同，但人们的反应却大不相同。

需要强调的是，人们对辛格的功利主义结论进行反驳时援引的很多因素在我们的研究中都得到了控制。与辛格的落水儿童例子不同，我们提供的所有场景中都不包含唯一能够提供帮助的人。同时，所有场景中人们提供帮助的方式完全相同，即通过声誉良好的组织进行捐款。需要帮助的事件究竟是紧急情况下的突发事件（挽救落水儿童）还是一直存在的问题（贫困问题），我们的试验也对此做出规定。为了排除由于爱国感情而做出的不公平判断，我们将所有场景都设置在国外。此外，有些不幸事件是由他人的行为所致，有人会因此认为导致事件发生的人应当负有更大的责任，我们便因此得以解脱。为了避免这种想法，我们设定的不幸事件都是意外情况。总之，我们所提出的场景中，除去空间距离的远近，几乎没有其他不同。由此可以得出结论，道德义务感深受空间距离及其他类似因素的影响。*

单纯的空间距离重要吗？在小火车问题中，空间距离确实会影响我们对某人性格的评价。如果某人因为担心自己的套装而任凭一个孩子在自己面前溺水而亡，他便是道德上的禽兽。但如果把钱花在购买套装等方面，而没有用于捐赠，则不能算作禽兽之举。然而这并不能证明空间距离的重要性，只能说明对空间距离不敏感的人在道德上有异于常人，而这种异于常人的状态十分重要。假设一位朋友给你打电话，就道德问题寻求建议："我应该帮助这些可怜的台风受害者吗？"如果你说："嗯，这要视情况而定。你距离他们有多少英尺？"这样的回答非常奇怪。

我们似乎再次被僵化刻板的自动模式误导：身旁的落水儿童触动了我们的道德神经，而远方的饥饿儿童则没有，*但两种情况的区别其实就像个人力量的介入一样，与道德并不相关。道德神经为什么会是这样？或者我们应该问：道德神经为什么不是这样？据我们所知，同理心最初是为了促进合作而产生的。得到促进的并非普遍合作，而是与某些特定的人或是与某个族群内部其他成员间的合作。如果你向一位急需帮助的族群内成员伸出了援手，那么在未来的某个时间，这位"身处患难的朋友"会变成你的"患难之交"（互惠原则）。帮助部落内部其他成员的同时，你也使整个部落变得更加强大，胜过了与其竞争的其他部落。通过这种方式，你也间接地为自己提供了帮助。（今天你救起的落水儿童可能会在未来某天带领你的部落进行战斗。）相比之下，从生物学角度来看，滥用同理心并不能带来任何优势，这一结论仅限于生物学层面。某些特点之所以能在自然选择中胜出，是因为它们能带来个人层面或是群体层面的竞争优势。我们往往不会被在远方受苦的人们感动，便是出于这样的原因。从生物学角度来看，更难解决的问题是，为什么我们有时会被身旁受苦的陌生人感动？站在文化进化的角度也许更易回答这个问题。正如第3章所述，根据某些文化规范，陌生人应当完全无私地相互对待，至少在代价不太高的情况下应当如此。

身旁的落水儿童与远方的贫困儿童相比，还有另一个显著区别。落水儿童是具体的、可辨认的个体。但从你的角度来看，通过捐款挽救的儿童是不明身份的、"以数字代表的"个体。*经济学家托马斯·谢林（Thomas schelling）观察到，与不确定的、"以数字代表的"受难者相比，人们对可辨认个体的需求会做出更加迅速急切的反应。这种现象被称为"可识别受害者效应"，杰西卡·麦克卢尔（Jessica McClure）的事例便反映了这种效应。

1987年，18个月大的婴儿杰西卡在得克萨斯州米德兰市落入井中，被困将近60个小时。陌生人为她的家庭捐赠了70多万美元，用于救援行动。这笔钱如果用在预防性医疗措施上，足以挽救很多孩子的生命。任由杰西卡死在井中简直是无法想象的事情，是道德上的禽兽之举；但不增加国家在

儿童预防性医疗措施方面的预算似乎并非如此难以接受。谢林在可识别受害者效应研究的开创性论文中指出，某个特定个体的死亡会唤起“焦虑和感伤、内疚和畏惧、责任与宗教等情感，（但）……当我们面临以数字代表的死亡时，这些敬畏便消失不见了”。

受到谢林的启发，黛博拉·斯莫尔（Deborah small）与乔治·卢文斯坦（George loewenstein）进行了一系列试验，检验人们对可识别的受害者与“以数字代表的”受害者的态度有何不同。首先，10位参试者每人会获得10美元作为“基金”，随后他们要随机抽取一张卡片。如果抽到了写有“保留”的卡片，参试者就能将基金留下；如果抽到了写有“失去”的卡片，基金就会被拿走，该参试者便成为“受害者”。随后，非受害者与受害者抽取号码，两两分为一组。重要的是，非受害者并不知道与他们同组的受害者身份。作为非受害者，你只知道“4号”与你同组，但你不知道，也无法知道“4号”是谁。非受害者可以将一部分自己的基金送给同组的受害者，赠送的金额由非受害者自己决定。此过程中最重要的控制因素是，有些非受害人是在确定分组后（两人并未见面）才决定给特定的受害人赠送多少基金；而有些非受害人则被告知，他们要在分组之前决定赠送的金额。所以有些人面临的问题是：“我要给4号（确定的受害者）多少钱？”其他人面临的问题则是：“一会儿通过抽签分组，我要给同组的那个人（不确定的受害者）多少钱？”在本试验中，决策者完全无法知道受赠人的身份。

这项试验中，确定的受害者得到的平均金额比不确定的受害者得到的两倍还多。也就是说，与“不明身份者”相比，人们更愿意把钱送给“随机挑选的4号”，这完全说不通：既然无论如何都无法了解赠送对象，那么先选择接受对象还是先选择赠送金额显然是无所谓的。

在后续的研究中，斯莫尔和卢文斯坦让人们估测自己的同情心（与“同理心”概念相同）程度，结果显示，对受害者的同情心程度与随后的赠送金额直接相关，这与人们所预想的结果相符合。在一项现场试验中，试验人员为参试者提供机会，向国际仁人家园捐款，每笔捐赠都会用于为有需要的家庭提供房屋。对某些人来说，接受捐赠的家庭已经提前确定；对其他人来说，接受捐赠的家庭则尚未确定。决策者对接受捐赠家庭的情况依然一无所知。同试验人员的预测一样，提前确定的家庭接收到的捐款几乎是未确定家庭得到捐款的两倍。近期一项研究显示，同我们所想象的一样，接受捐赠者如果是某个具体的需要帮助的人，比如一个贫穷的7岁的马里女孩儿洛基亚，人们便更愿意为慈善掏腰包。但如果慈善捐款是为了解决非洲贫困等更大的问题，人们的捐赠意愿会变弱。我们还预想到，人们对洛基亚的捐款数额与自己估计的同情心程度密切相关。但出乎意料的是，如果研究人员将非洲贫困问题的相关数据和洛基亚的个人故事同时给出，人们帮助洛基亚的意愿也减弱了，因为这种情况下，洛基亚的经历成

了“大海里的一滴水”。事实上，同情心的减弱并不需要惊人的巨额数字。特希拉·科格特（Tehila Kogut）和伊拉娜·里托夫（ilana ritov）向人们募捐时，一边是一位需要巨额医疗费用的患病儿童，另一边是8名情况类似的儿童。人们对一位儿童显示出了更多的同情心，与其他8名儿童相比，这名患病儿童得到了更多的捐款。近期一项研究表明，数字上升到2的时候，我们的同情心便会大打折扣。*

所以……身旁的落水儿童与远方需要食品和药物的儿童之间，真的存在什么道德区别吗？这些事例给人的感觉当然不同。但现在我们得知，对于道德义务的直觉判断有时并不可靠，我们会对不重要的因素十分敏感，比如空间距离，比如是否能够知道被帮助对象的哪怕一丁点儿的信息。这并不是说负责同理心的大脑机制不好，相反，如果失去了发自内心的同理心，我们就会成为道德禽兽。人类的同理心也许是大脑中最典型的道德特点。但无论怎样，如果将同理心机制这种刻板的判断方式作为基本的道德准则，那就太过愚蠢了。

## 个人承诺

姑且认为，我们对远方的、“以数字代表的”人们所遭受的苦难表现出了不应有的冷漠。那么，帮助远方陌生人的责任是否应当高于一切？我们所关心的其他事情该怎么办呢？

人类不仅是资源的分配者，也同时扮演着父母、子女、手足、爱人、朋友、公民的角色。我们还是自身信念的守护者，无价事业的开拓者，不论是对艺术、对知识的追求，还是对美好生活的向往。这些承诺让我们背负着合法的道德义务和道德选择。如果你把钱捐给了远方某个贫穷的、不知名的孩子，从来不给自己的孩子买生日礼物，从某个角度来说你也许令人钦佩，但你绝对不是一位好家长。如果对功利主义思想的承诺让你失去了经济能力，进行交际活动，那么你绝对不是一位好朋友。同样，支持艺术或支持当地高中的体育团队似乎不能算是道德错误。追求全球幸福的最大化就一定意味着牺牲生活中其他有价值的事情吗？

关于这一点，功利主义思想能够做出很多协调。如果让一个真正的人为了不知姓名的陌生人抛弃家庭、朋友和其他激情这个想法听上去十分荒唐，那么这便不可能是功利主义对真正的人所提出的要求。如果尝试这样做，就会变成一场灾难，而灾难并不能将幸福感最大化。人类进化过程中追求的生活以人和社区的关系为基础，如果我们的目标是让这个世界尽可能幸福，那么就必须考虑到人性中这个决定性的特点。

除了温和的功利主义协调之外，我们还需要一些挑战性的改革。给孩子买生日礼物毫无疑问是没错的，但你的孩子真的需要3份生日礼物吗？或者5

份？10份？从某种程度来说，将钱花在自己的孩子身上而不是捐给那些急需食物和药品的孩子，这种行为在道德上也许真的不对。支持艺术是件好事，也许比把钱花在自己身上要好一些。但将100万美元捐赠给美国大都会博物馆似乎在道德上站不住脚，这笔钱足以购买一件价格适中的世界级艺术品，足以为1000名贫困儿童提供食物、医疗、衣物和教育。在实际情况中，对出于好心但方法不对的慈善家们嗤之以鼻并无助于情况的改善。当然，把钱捐给大都会博物馆总好过为自己购买第4套度假屋，但对需要帮助的人伸出援手依然不失为更好的选择。在个人私欲和人际关系方面，当涉及崇高的事业时，对已有资源的合理使用和铺张浪费之间并没有清晰的界限。但在现实世界中，我们不能把界限画在显而易见的荒唐之处，否则没有人会尊重这条界线。

可以看出，功利主义一如既往的坚定合理，在实践中对我们的需要与局限进行协调。但我们依然感觉有一些深层次的重要人类价值似乎被忽略了。

## 人类价值与理想价值

功利主义思想允许我们经营自己的人际关系和兴趣爱好，能够对我们表示宽恕。但我们有必要为这些事求得宽恕吗？你会说，道德上的完人并不是幸福之源，在朋友和家人身上投资也并非差强人意的人性弱点，能这样做的人是朴素简单的好人。如果功利主义的理想是将陌生人的不幸置于其他一切事情之上，那么这种思想是不是不太对劲儿呢？

也许不是。如果我们从人类价值中抽身出来，站在足够远的距离外回头审视，便会发现，即使我们依然决定拥护人类价值，这种价值也不是理想价值。下面的思维试验也许能够帮助我们理解这一点。

假设你是宇宙之主宰，决定创造一种智能的、有感情的新物种。这个物种和我们一样，生活在资源有限的世界中，将资源分配给穷人比分配给富人更能够缓解人们的痛苦，产生更多的幸福。你需要为新物种设计一种思维方式，决定他们彼此相待的方式。你的选择范围已经缩小到了三个物种：

> 物种1 自私类人物种：这种生物对彼此完全不关心。只要能让自己尽可能开心，他们什么都可以做，完全不关心他人的幸福。自私类人的世界十分悲惨，毫无信任可言，每个人都在不断争斗，抢夺稀缺资源。
>
> 物种2 完全类人物种：这种生物十分自私，但他们对一小部分特定人群的利益也十分关注，同时对属于某个群体的个体也有一定程度的关心。如果其他因素保持不变，他们愿意看到他人幸福，不愿看到他人

受苦。多数情况下，他们对陌生人的帮助仅限于举手之劳，特别是对于其他群体的陌生人，他们似乎更加不愿付出。这是一个有爱的物种，但他们的爱范围不大。这种生物中很多个体都很幸福，但整个物种却不像其应有的那样幸福。这是因为完全类人会为自己和最亲近的人积累尽可能多的资源，导致很多完全类人（比一半略少）得不到达到幸福所需的资源。

物种3 功利类人物种：这种生物将所有成员的幸福一视同仁。他们的生活极其幸福，因为其成员关心别人的程度毫不亚于关心自己。这种生物充满了博爱的精神，他们彼此相爱的程度与完全类人对自己的家人和密友的感情不相上下。因此，这是一群非常快乐的生物。

如果我是宇宙的主宰，我就会选择功利类人这个充满博爱的快乐物种。你可能会有不同意见，坚持认为功利类人成员像是没有头脑的蜂群，他们不加选择的博爱在人类丰富且有针对性的情感面前黯然失色，就像是博格人和罗密欧与朱丽叶两者的对比。但持有这种观点只能说明你的想象力还不够丰富。为了帮助你展开想象，请回想一些现实生活中的英雄。有些人将肾脏捐赠给陌生人，却不求任何回报。更令人惊叹的是，这些人并不认为自己是英雄，他们心中怀有暖人的乐观精神，坚持认为如果别人有机会，也会做出同样的选择。还有韦斯利·奥特利，他为了拯救一名癫痫发作摔倒在轨道上的男子，不顾前方驶来的地铁，跃进轨道，将那名男子护在身体下面，火车擦着奥特利的头发，从两人上方呼啸而过。功利类人让我们联想到的形象是英雄，而不是蜂群。他们是和我们一样的人，但他们愿意为他人所付出的，比我们多数人都要多。

我想表达的是，要求人们在现实生活中抛开自己热爱的一切，只为实现更大范围的利益是不合理的。至于我自己，我把钱花在了子女身上，没有帮助远方的饥饿儿童，我也没有改变现状的想法。不管怎么说，我只是个普通人！但我宁愿做一名普通人，认识到自己的虚伪并努力改进，也不愿将人类特有的道德局限当作理想的价值。

## 公正的荒漠

功利主义要求对打破规则的人施以惩戒，一个直接的原因是：如果没有惩罚的威胁，人们便不会遵守规则。但有人认为，惩罚的首要目的不是，也不应是对正确行为的鼓励。他们认为，之所以要对犯规者进行惩罚，是因为他们应受惩罚，而不应考虑惩罚所带来的实际好处。这种惩罚方式被称为报应主义，包括伊曼努尔·康德（immanuel Kant）在内的很多道德理论家与法律理论家都对这种观点情有独钟。事实上，康德曾说过，如果某个岛国要集体离开家园，那么岛民们离开前要做的事情清单中应当包括处决

监狱里所有的谋杀犯，这样是为了在离开前再多实现一点公正。

报应主义的拥护者对功利主义提出了很多有力的反驳。首先，功利主义有时似乎会在不恰当的时候实施惩罚。可能你还记得第3章中治安法官与暴民的例子，治安法官可以通过监禁一名无辜平民来避免一场暴动。即使这种做法能够导致更好的整体结果，惩罚一名无辜平民显然也是错误的做法。其次，功利主义的惩罚有时似乎太过轻微。对于报应主义者来说，理想的世界应当善有善报，恶有恶报。但对功利主义者来说，理想的世界中，每个人都应达到最大限度的幸福，包括坏人在内。事实上，在理想的功利主义惩罚机制中，惩罚是一种有效的威胁，并不必真正落实。在理想的功利主义世界里，罪犯会被送到一个快乐的地方，他们在那里不会对任何人造成妨碍，其余的人们则相信罪犯遭到了应有的惩罚，大家会因此更加守规矩。

无辜者受罚？有罪者得奖？功利主义似乎对公正的声音充耳不闻。有人认为我们完全有理由排斥功利主义理想，拒绝将幸福感最大化。与之前一样，我们要牢记现实世界的规则，尽可能多地将功利主义与人们的常识相协调。

从政策层面来看，关于惩罚无辜者、奖励有罪者的担忧没有任何现实依据。我们可以虚构出无辜者受罚导致事情向好的情景，比如治安法官和暴民的例子，但在现实生活中，这样的政策糟糕透顶。虚假惩罚的例子也是如此，类似的政策要想达到功利主义目的，政府官员就必须与奥威尔式的另一部分官员结成同盟，制定无限期的密约，还要每天抵制滥用权力的诱惑。这样不可能让世界更加幸福。

功利主义能够自然地协调其他常识性正义之间的矛盾。例如，如果人们对他人造成的伤害属于无心之失，所受到的惩罚就会大大减轻（甚至免于受罚）。上一章曾提到，功利主义可以完美地解释这条常识性规则：总体上讲，故意造成伤害的人远比无心造成伤害的人危险，因此对前者的威慑非常重要。此外，蓄意的行为受到人的意识控制，因此惩罚的威慑也更有可能使人放弃这样的行为。当然，如果人们因疏于职守而造成伤害，我们也会对这种无心之失施以惩戒。这种做法也有其功利主义依据：我们既要防止有意造成的伤害，也要阻止危险的疏忽。

同样的，法律和常识所认可的合理借口和原因也能从功利主义角度得到解释。例如，法律规定“未成年”（孩子的身份）是一种合法的例外，功利主义也可以对此做出解释：与30岁的成年人相比，10岁的孩子犯罪后更容易在温和手段的介入下改过自新；面对严酷的惩罚，也更可能遭受不可逆的损害。精神不正常的人对惩罚的威慑作用不敏感，因此从功利主义角度看，对其施以惩戒的理由便不那么充分。此外，对于因正当防卫行为

和“紧急避险”情况（例如，为了救人或者救己而偷窃船只）违反法律的行为，功利主义也有其自然的解释：我们并不想阻止人们的类似行为。

因此，在现实世界中，以最大化幸福为目标的法律体系中并不会出现危险的奥威尔式规则，无辜的人不会无端受罚，有罪的人也不会无故被宽恕甚至得到奖励。在现实世界中促进更大范围利益的惩罚体制也会含有各种各样的合理借口和原因，将蓄意犯罪和无心之失区分开来，将孩子和成年人区分开来。这也就意味着，以更大范围的利益为唯一目标的惩罚必然会包含争议性的改革。

例如，对监狱中犯人的安全和幸福的考虑几乎无法得到公众支持，对政治家也是一种负担。有些犯人经常遭到性虐待，至少有些人会对此感到遗憾，但这件事不会对我们造成过多困扰，我们也不会为狱中的受害者争取更多的保护。但我们需要考虑这个问题：如果一项政策规定，作为惩罚的一部分，犯人会被正式强奸，你会表示支持吗？监狱中的强奸是可预见的连带反应，其受害者只能听天由命，很多人对这种情况表示震惊和遗憾，但也表示可以容忍。如果将强奸作为国家刑罚的手段，由某些特定的个体刻意执行，这便是野蛮的行径。然而，从功利主义角度来说，两种性虐待的方式在道德层面的区别并不很大。强奸就是强奸，我们应当尽量避免犯人之间的此类暴力行为。

在刑法方面，减少犯人之间的性暴力只是一个例子，反映了感觉正确的事和对社会真正有利的事两者间的冲突。关于监狱生活的本质和后果，还有一个更加抽象的问题：犯人出狱以后，狱中的生活会让他从此以后更加勤奋守法吗？狱中的悲惨生活会让其他人更加安分守己吗？毫无疑问，总体来说，实施惩罚确实是一种重要的威胁，能够有效阻止犯罪。但对美国社会所特有的社会问题来说，严酷且频繁的惩罚是必要的应对手段吗？或者这种惩罚只是不合理的政策？有犯罪预谋的人了解当地的法律法规吗？他们在乎吗？我的整体观点是，以更大范围的利益为目标的刑事审判体系不会沦为荒谬的奥威尔式机器，但与高度体现报应主义思想的现有刑事审判体系相比，可能也会有很大的不同。

## 理想中的公正

功利主义的公正在实践层面是合理的，但在永恒的“理论层面”，一些问题还有待解决。假设对无辜的人施以惩罚确实能促进更大范围的利益，那么这种做法是正确的吗？假设我们真的能够以较低成本实现有力的惩罚威慑，那么让谋杀犯和强奸犯免于受罚，过上舒适的生活，真的是更好的选择吗？（假设这样做并无损于惩罚措施的威慑作用等积极影响）对于功利主义者来说，惩罚是必要的邪恶，但对做了坏事的人施以惩罚，这种行为

本身难道一点可取之处都没有吗？在实践层面，不论功利主义思想中公正的概念多么正常，一旦涉及公正的深层含义，功利主义似乎遗漏了什么。

这就是批评家们的观点。另一种可能性是，直觉中的正义感由一系列启发性的思想构成，是一种非常有用的道德机制，但并非绝对可靠。我们对惩罚有一种偏好，同所有偏好一样，对惩罚的偏好微妙且复杂，受到遗传、文化、特殊品质等一系列复杂因素的影响。它由大脑的自动模式执行，也因此在灵活性方面有很大局限。所有的偏好都可以被愚弄：人造甜味剂可以愚弄味蕾对甜味的偏好；避孕措施和色情产品可以愚弄人们的性欲，因为两者都能提供性快感却无涉于基因的传播。有些时候，我们也会被偏好所愚弄：对脂肪和糖分的偏好让我们在物质丰盈的世界里变得肥胖；毒品的滥用绑架了我们的奖赏回路，摧毁了人们的生活。要想知道究竟是谁愚弄了谁，我们便需要冲破偏好的局限，站在新的视角考虑：减肥汽水、色情片、能多益、海洛因等事物究竟在多大程度上符合我们的最终利益？对于惩罚的偏好，我们也应提出同样的问题。

如前所述，人类直觉中的公正感非常重要，如果没有公正感，我们便会迷失自我。第2章提到，惩罚能够鼓励人们做有利于“我们”而不是“我”的事，从而促进合作。也就是说，惩罚的自然功能与功利主义类似：我们是天生的惩罚者，因为惩罚本身就具有一定的社会功能。*

如果你向人们询问，为何要对违规者进行惩罚，人们给出的答案带有明显的功利主义痕迹：如果没有惩罚的威胁，人们便会胡作非为。这是手动模式给出的答案，但在某些特定的案例面前，人们做出的惩罚判断明显表明，他们首先考虑的绝对不是威慑。第2章中提到，惩罚大多源于气愤、厌恶等感情，这些感情并非出于阻止未来违规行为的考虑而产生，而是因违规行为或是违规者本身而引起的。人们考虑对违规者的惩罚方式时，往往会忽略与震慑力相关的因素，仅仅根据自身对违规行为的感受来决定惩罚方式。比如，对于难以侦查到的罪行施以更重的惩罚，这种行为符合功利主义思想，因为如果罪行被抓到的风险很低，就需要有其他的震慑因素。（比如，在加利福尼亚州，针对乱丢垃圾的罚款额度高达1000美元，这并非因为往地上扔纸杯会造成多么可怕的破坏，而是因为乱丢垃圾后侥幸逃脱的可能性很大。）人们往往会忽略功利主义在这方面的考量。破案率较低的罪行并不会让我们更加生气，因此我们不会出自本能地施以更重的刑罚。第2章提到，人们常常忽视惩罚成本和收益之间的关系，这也许是惩罚的一个基本特点：如果只在“值得”的时候才惩罚，那你便不是可靠的惩罚者，别人会轻易将你作为突破口。但如果你性情鲁莽，复仇心强，并且以此为他人所知，那么你的威慑便更加有力。

某些情况下，我们做出的惩罚判断十分不合理。斯莫尔和卢文斯坦曾经记录人们帮助特定受害人时所表现出的偏好，他们在另一个关于惩罚的试验

中也记下了相似的行为。参试者被要求玩一个游戏，既能够合作完成，也能够以利己为原则完成。游戏之后，试验人员给选择合作的玩家提供一个机会，惩罚那些以自我为中心的玩家。有些合作者可以匿名惩罚某一个特定的人："你希望怎样惩罚自私的4号？"有些人则可以匿名惩罚某个不确定的人："你将从自私的玩家中随机抽取一个号码，你希望怎样惩罚他？"不出意料，人们对"确定的"违规者施以的惩罚力度几乎是"不确定"玩家的两倍，人们给出的惩罚方式与他们的情绪反应也是成比例的。

追究道德责任时，情绪也会影响我们的判断。肖恩·尼科尔斯（shaun Nichols）和乔舒亚·诺布（Joshua Knobe）向人们描述了一个"决定论"的世界：

想象一个世界（A世界），每件事情都完全由之前发生的事情导致。这个世界从一开始就是这样，最初发生的事导致了接下来的事，一直这样下来，直到现在。比如有一天，约翰决定中午要吃炸薯条，与其他所有事件一样，这个决定完全建立在前一事件的基础上。所以如果在此之前，这个世界上发生的事情完全不变，那么约翰吃薯条的决定是必然会发生的。

尼科尔斯和诺布让参试者考虑，这个世界的人在道德上对自己的行为是否应负完全的责任？只有不到5%的参试者给出了肯定的答案。另外一组参试者也阅读了上面一段关于A世界的描述，但他们需要回答的问题并不是关于责任的抽象问题，而是更加具体的、能够将情绪调动起来的问题。

在A世界中，一个名叫比尔的人爱上了自己的秘书，他认为与她在一起的唯一方式就是杀掉自己的妻子和三个孩子。他知道一旦遇到大火，任何人都无法从家中逃生，于是在出差之前，他在家中的地下室里安了一个装置，这个装置将会烧毁房子，杀死他的家人。

这一次，72%的参试者认为，从道德上讲，比尔对自己的行为应当负全责。这种态度的彻底转变实在令人吃惊。在决定论的世界里，以抽象的方式向人询问责任问题时，几乎所有人都认为道德责任并不存在；但面对具体的不法事例，情感对人们的判断产生了影响，之前的抽象判断便飞到了九霄云外。

康德认为，惩罚违规者的行为本身便具有一定的价值。如果这种说法是正确的，这便是一个惊人的巧合。自然选择将惩罚机制放入人类大脑，从而促进合作，更好地传播人类基因。但如果真正的公正原则与惩罚机制的感受刚好相同，这该是多么奇怪的情况。基于对大脑的进化过程及其工作机理的了解，我们可以做出合理的假设：人类对公正的偏好也许只是一种有用的错觉。我们认为惩罚不仅是使人规矩行事的手段，其本身也是有价值的；就像我们认为食物不仅是获取营养的手段，其本身也是美味的。我们

从食物中获得的快感完全无害，但对他人施以惩戒却不可能完全无害。因此，对于大快人心但弊大于利的惩罚，我们应当保持警惕。功利主义者的视线超出了我们对惩罚的偏好，看到了更远的地方，我们不应为此责怪他们。

## 公正的社会

功利主义是一种讲求平等的哲学思想，要求富人为穷人做出很多贡献。如果你明天一早醒来，思想获得重生，变成了一名功利主义者，那么你生活中最大的改变莫过于对帮助不幸者产生了极大的热情。尽管如此，功利主义长久以来遭人诟病的一点依然是不够平等。人们认为功利主义（可能）无法尊重弱势群体的利益。

约翰·罗尔斯（John rawls）是20世纪最重要的道德哲学家，他认为将幸福最大化的行为可能会导致极端不公。功利主义认为，如果降低一部分人的幸福感有助于使其他人获得更大的幸福感，这种做法是可以接受的。这便是累进税制的理论基础：让富人缴纳更多税金并不会扰乱他们的生活，但这笔税收可以为社会带来很多好处。尽管如此，罗尔斯认为，以最大化幸福感为目的分配资源可能会造成不公正的现象。回想第8章中罗尔斯提出的例子：假设一个社会中，多数人将少数人作为奴隶，如果多数人对这种安排很满意，其幸福程度也足以抵消被奴役者的不幸福，这种做法合理吗？罗尔斯认为，有序社会的基础应当是某些基本权利和自由，而不是压倒一切的将幸福感最大化的目标。

从表面看来，这个观点十分有力。毫无疑问，奴隶制度是不公正的，支持奴隶制度的道德标准也是不合理的。但问题在于，功利主义是否真的支持奴隶制？为了解决这个问题，我们需要把问题分为两部分："理论上"和"实际上"。我将重点解决"实际上"的问题，因为对于本书观点来说，"实际上"的问题才最为重要。我的观点并非声称功利主义是绝对的道德真理，我想表达的是，功利主义是一种恰当的元道德，为现实世界中的道德争执提供了一种好的解决标准。只要功利主义不在现实世界中支持奴隶制，那就足够了。

在现实世界中，我不相信将幸福感最大化的目标会导致奴隶制等事件发生。作为一名坚定的经验主义者，我极其不愿坐在扶手椅里对外面世界的情况妄加评论，但在这件事上，我要大胆地说，功利主义在理论上也许支持奴隶制，除非人性发生巨变，这种改变不可能发生。因追求幸福感最大化而导致奴隶制的事件只有在科幻小说里才会发生。（在科幻的世界里，人类的道德直觉不一定可信。）

要将功利主义和社会公正的关系理清，是一件异常困难的事。具体来说，

由于我们常常自然地将功用与财富混为一谈，“功用”的概念变得很难厘清。我们会在后文谈论“财富主义”的谬误，但现在，我想以一种不同的方式证明，奴隶制以及其他形式的压迫很难让世界幸福起来。

奴隶制已经为某些人创造出巨额的财富，也造就了无数的痛苦。当我们抽象地综合考虑收获和损失时，实际情况中的损失未必大于收获，这种考量看上去更像是开放性的经验主义问题。但我不这样认为，要想更加清晰地考量奴隶制对人类幸福感的影响，我们需要拉近镜头，观察某些代表性个体的幸福感受。

奴隶制社会中当然会包括奴隶和奴隶主。为了进行更加具体的讨论，我们设想一个典型的奴隶制社会，其中一半人是奴隶主，一半人是奴隶。也就是说，每个自由人都刚好拥有一个奴隶。（请注意，与稍后提出的观点相比，这种一对一的比例属于保守假设。**）如果奴隶制能够将幸福感最大化，那么每个奴隶主因拥有奴隶而获得的平均幸福感必须大于他的奴隶因被奴役而失去的幸福感。这种说法合理吗？

让我们来分阶段考虑这个问题。假设你既不是奴隶也不是奴隶主，第一个问题是：如果拥有一个奴隶，你的幸福感会增加多少？当然，因为你是一个好人，拥有奴隶这件事也许根本不会让你更加幸福。但我们只是试着想象一名奴隶主的生活会有多幸福，暂时没有道德禁区。为了简化问题，可以假设你有一个高科技的机器人奴隶，一个身体健壮但没有文化的人能做的事，机器人奴隶都能做，但机器人奴隶与你的笔记本电脑和烤箱一样，没有任何感情。这样一来，拥有机器人奴隶的你便和过去拥有真正奴隶的奴隶主一样，不会因奴隶的问题感到困扰了。

你会如何对待机器人奴隶呢？按照过去大多数奴隶主的做法，他们会从奴隶身上榨取尽可能多的经济价值。你可能会让奴隶去工作，假设你的奴隶每年能为你额外挣得5万美元。（这也是一个保守假设，对拥有奴隶的好处做出了过高的估计。即使经常加班，一位技术不熟练的工人一年也很难创造出5万美元的价值，况且我们尚未计算为了满足奴隶的基本需求所花费的成本。）一年多挣5万美元的感觉怎么样呢？还不错，但也许没有你想象的那么好。如果你拥有一个奴隶，说明你的经济状况已经相当不错。从诸多关于幸福感的研究中，我们学到的最基本的一点就是，额外的收入（适度收入水平以上）对一个人幸福感的加成很小。*有些研究表明，收入水平达到一定高度后，额外的收入甚至不会增加任何幸福感。每个人因额外收入而获得的幸福感各有不同，但我们知道的是，从平均水平看，一定数量的额外收入对富人的影响很小，对穷人的影响则很大，这才是真正重要的。需要强调的是，这项结论并非尝试性试验的结果。*几十年的研究发现，财富（适度收入水平以上）与幸福之间的联系很弱，这是一条揭示人类本性的规律。超过一定水平之后，财富无法买到（很多）幸福。

因此，我们可以保守地得出结论：拥有一个奴隶会让你在财富方面收获颇丰，幸福感也有适当的增加。现在你需要考虑的是第二个问题：如果你变成一名奴隶，你的幸福感会减少多少？答案当然是减少很多，原因也是显而易见。我不会浪费过多的笔墨控诉奴隶制的恐怖。历史上对奴隶的虐待包括殴打、强奸、做苦力等，奴隶没有任何人身自由，家庭常常离散。即使是在最好的情况下，变成他人财产的一部分也不是一件好事，在某些情况下，这甚至是无法想象的人间惨剧。如果你从今天的自由之身变成一名奴隶，无须多想，你的幸福感必然会直线下降。

考虑过个体因奴隶制而收获或损失的幸福之后，你可以回答第三个，也是最后一个问题了：奴隶制所带来的收获真的大于损失吗？也许考虑这个问题的最好方式是将其转化为对等的个人选择问题：一生之中，你是否愿意花一半的时间作为奴隶，以交换在另外的一半时间里每年多挣5万美元？还是你宁愿保持现在的生活方式不变？我希望你的答案是显而易见的。如果这个问题的答案显而易见，那么经过具体化的思考，同样显而易见的是，奴隶制在现实世界完全不能将幸福感最大化。同样的道理可以解释广义上的一切压迫行为。总体不公是总体上的不公正，因为某些人承受的后果极其悲惨。在某些人的想象中，“功用怪兽”能够通过吃人获得无限的快感，但回到现实中，即使是再好的人，我们从其身上获得的好处也抵不过对于压迫的恐惧。**

现实世界中确实存在的，是能够促进更大范围利益的社会不平等以及对自由的限制。自由市场政策会导致经济上的不平等，使人们开始思考如何对财富进行重新分配，或者是否应当对财富进行重新分配。在对财富进行最大限度重新分配的社会中，虽然消除了不平等，但一同消失的还有人们积极生产的经济动力。理解了这一点，几乎所有人都会像北方牧民一样，为了提高生产效率（如果不考虑公平的因素），接受一定程度的经济不平等。自由权利的不平等也是同样的。美国和其他一些国家规定，如果某人的HiV（Human immunode？ciency Virus，人类免疫缺陷病毒）检测呈阳性，在性伴侣不知情且不采取任何保护措施的情况下进行性行为是违法的。这类法律限制了HiV阳性公民的自由，而HiV阳性人群属于弱势群体，但大多数人依然认为，为了更大范围的利益，这类法律是合理的。对更大范围利益的考量在其他方面也限制了我们的自由，比如在拥挤的剧院大喊“着火了”这个老生常谈的例子。结论：我们确实可能为了更大范围的利益而造成不平等，限制人们的自由，但在现实世界中，这种不平等和限制绝对不会导致总体不公。有些事对某些人来说也许不太公平，但问题在于，现实世界中的功利主义绝对不会像批评者所描绘的那样，导致奴隶制等明显不平等的社会安排。在理论层面你可以尽情诟病功利主义，但在实践当中，让世界尽可能幸福的行为并不一定会导致压迫。

既然如此，为什么有如此多的智者得出结论，认为功利主义必然导致总体不公呢？如前所述，一部分原因要归于人们对“功用”这个词的误解。人们往往误把功用与财富等同，这样一来，将功用最大化的概念便失去了吸引力，甚至有些不公平。乔纳森·伯龙（Jonathan baron）和我在一项试验中记下了这种混淆，后文将会对此详细介绍。这项试验有些复杂，如果你想要跳过的话就请便吧。对于选择跳过的读者来说，你们需要了解：如果你认为压迫在现实世界中能够将幸福最大化，那么你的想象是不正确的。你的观点是，能够通过压迫得到最大化的是财富，而不是幸福。

## “财富主义”谬误

如前所述，“功利主义”是绝妙想法头上的糟糕名字。我们由“功用”所联想到的意象并不符合功利主义思想。

第7章曾经提到，功利主义的第一个核心思想是体验的重要性：好的事情之所以好，坏的事情之所以坏，都由它们对人类体验的影响而决定。（再次强调，功利主义的第二个要点是，每个人的体验都同等重要。）体验是衡量功利主义的货币，既包括广义上幸福的体验，也包括不幸的体验。但“功用”给人的感觉更像是“有用的东西”，拥有很多有用的东西，就是拥有财富。因此，功利主义很容易被错认为“财富主义”，把拥有最多财富作为首要目标，这显然不是一个好观点。

然而人们对功利主义和财富主义的混淆远不止一个词语的理解。“功用”、“幸福”或“体验的质量”等概念都很难被人正确理解。我们已经习惯将事物量化，量化社会上的事件、量化事物的特点：苹果有多少？水有多少？会议开多长时间？多少平方英尺？钱有多少？但通常情况下，我们不会将体验的质量进行量化。因此，当我们对“功用”进行分配时，便会很自然地联想到分配事物，而不是分配体验的质量。

功用与事物十分相似，但其本身并不是事物。首先，功用不一定来自市场上的商品。友谊、晴天、证明数学定理、受到邻居的尊敬都能使你获得积极的体验，这就是“功用”。其次，功用并不等同于事物。因为对于一定量的事物来说，不同的人在不同情境下获得的功用也各不相同。对一位贫穷的柬埔寨农民来说，2000美元可以轻易地改变他的生活；但对一位富有的商人来说，2000美元不过是飞往新加坡的旅途中将座位升至头等舱的一笔费用。因此，对功用的分配进行讨论时，我们需要牢记，被分配的是功用，是一种难以捉摸的、精神层面的东西，而不是社会事件或银行账户里的数字。

那么我们的问题就是，将功用（不是事物，也不是财富）最大化过程中产生的社会不平等是否会导致现实世界中的总体不公？罗尔斯设想功利主义

可能造成的不平等时，想到了奴隶制。毫无疑问，奴隶制是不公正的。但为什么会有人认为奴隶制等类似的事情会使整个世界更加幸福呢？如果将功用与事物或财富混淆起来，奴隶制能够将功用最大化这个观点听上去就合情合理了。

如前所述，奴隶主因拥有奴隶而获得的幸福（功用）并没有超过奴隶因被奴役而损失的幸福。也就是说，你不愿以半生的奴隶生活换取另外半生的大富大贵。同理，2000美元对于富商和贫农的意义也不可能相同，但如果我们从财富而不是幸福的角度考虑这个问题，那么这种对富人有利、对穷人不利的计算就能说得通了。奴隶主可能会通过剥削奴隶获得巨额财富；副总统在旅途中度过舒适的一晚后可能会促成或毁掉一笔百万美元的交易。从经济角度看，富人相对丰盈的收获可能会超过穷人相对不起眼的损失，当罗尔斯和其他人考虑功利主义的不公正时，他们就是这样想的。快乐的剥削者会说："真抱歉，因为我的收获很多很多，因此你的损失虽然很多，但也是可以接受的。"然而，这种收获大于损失的结论是以美元计算得出，而不是以幸福来衡量的。**

之前提到的试验中，人们很容易将功用与财富相混淆，这便是罗尔斯所得结论的来源。在乔纳森·伯龙和我进行的试验中，我们为参试者提供了不同的虚拟社会模型，每个社会模型中人们的年收入（财富）分配状况都不相同。比如，在a国家中，底层1/3的人年收入为25000美元，中间1/3的人年收入为45000美元，而最富有的1/3的人年收入为70000美元。在b国家，底层1/3的人年收入仅为15000美元，其他的人收入水平不变。我们要求参试者据此比较a国家和b国家。试验中，我们假设参试者作为底层、中层与富有公民的概率完全相等，然后让他们根据自己是否愿意在该国生活，对该国进行打分。a、b两国的打分结果反映了人们对底层人民收入增加10000美元的态度，即这件事在人们心中的重要程度。同样的方式可以判断收入的增加在人们心中的重要程度，比如从40000美元到50000美元的增长。与你的设想一致，收入增加10000美元的效果不总相同。人们认为收入水平从15000美元增长到25000美元比从40000美元增长到50000美元更加重要，这反映了财富的边际效用递减。拥有的财富越多，额外1美元的意义就越微不足道，额外带来的幸福感也就越低。

在试验的下一个阶段，我们要求参试者评估第一阶段试验中各层次收入的价值，并进行打分。也就是说，参试者需要为每个收入层次分配一个功用等级。我们规定，最低收入打分为0，最高收入打分为100。参试者可以使用0～100之间的数字为其他的收入层次打分，但同样的分数差距所表示的价值应当相同。例如，分数从0上升到50带给人的愉悦感应当与分数从50上升到100相同。参试者的打分与我们的预期吻合，收入水平较低的区域，收入增加所得的分数更高，符合财富的边际效用递减规则。例如，收

入15000美元与收入25000美元的得分之差明显高于收入40000美元与收入50000美元的得分之差。

在试验的第三个阶段，参试者需要再次根据自己在各个国家生活的意愿打分，但这一次不再以收入水平来描述各个国家，而是通过功用分布进行描述。这一次我们不再提供每个社会的底层、中层，以及上层的收入情况，而是提供每个社会中三个阶层收入水平的得分情况（功用）。我们告诉参试者，这些得分情况是“像你们一样的参试者对收入水平的打分，就像你们在试验的上一阶段所做的一样”。我们由此可以看出数字的变化在人们心中的重要性，但这一次人们评估的并非收入变化，而是功用的变化。例如，我们可以判断出，从得分为0的收入水平变为得分为25的收入水平，这种变化在人们心中有多重要。我们也可以判断，从得分为75的收入水平变为得分为100的收入水平，人们认为有多重要。更关键的是，我们得以观察功用得分中不同阶段的25分对人们的吸引力是否也不相同。

如果人们的判断标准是稳定的，不同的25分对人们的吸引力应当相同，因为这次的考量因素是功用水平，而不是收入水平。参试者在试验的第二阶段已经打过分，功用水平轴上每一分所代表的价值都相等。也就是说，不论是从0分增加到25分还是从75分增加到100分，功用水平增加25分的效果应当相同。

结论：第二阶段试验中，人们为不同层次的收入水平分配功用；第三阶段试验中，人们依据不同国家的功用分布进行打分。但第三阶段的试验结果与人们在第二阶段中的表现不符，对于功用的数值增加，人们并不认为同样的数值代表同样的价值，而是采用第一阶段中对待收入水平变化的态度来考虑功用数值的变化。也就是说，他们与罗尔斯一样，认为数值较低区域的增量更加重要，从0到25的增量比从75到100的增量更加重要。事实上，考量功用分数时，人们对底层数值的重视程度与他们对收入水平的考量态度相同。这种评价模式本身是前后矛盾的，但这个结果刚好符合伯龙和我对人们行为的预测，这是我们阅读罗尔斯的著作后得出的结论。*

这项试验显示，人们很难对功用形成清晰的概念。一方面，人们能够理解事物与功用的不同。由于10000美元增量的基数不同（例如，从15000美元开始还是从40000美元开始），参试者给出了不同的分数，这便反映了这一点。另一方面，如果人们被要求衡量功用的分布，他们对待功用的态度与对待财富便没什么两样，并没有将功用当成特殊的抽象概念。也就是说，人们看到功用得分为0的收入增加到功用得分为25的收入时会想：“这个变化中，功用得分的起点那么低，其增量的影响一定很大。”然后，人们看到功用得分分别为75和100的收入，他们会想：“这是很大的改善，但其起点的功用得分已经很高了，那么这种增量也不能算是多大的进步了。”这种想法是完全错误的，在不同的情况下，从一定量的金钱当中获

得的功用也许有多有少，但从功用中获得的功用不可能有多有少。功用就是功用。这种情况之所以出现，并非因为参试者表现不好。当他们需要考虑的问题从收入分配转换为功用分配时，其思维并没有相应改变。他们将功用与事物以同种方式对待。（在我们的试验当中，很多哲学家也都以同样的方式思考问题。）

这意味着什么？这说明，经过事实的验证，罗尔斯对功利主义的批评以及认为功利主义会导致总体不公的观点都站不住脚，他对功利主义的驳斥源于对功用一词的误解，这一点已经轻易地在实验室中得到证实。人们在某种程度上能够理解财富与功用的区别，他们能够理解，随着财富的增加，增量的重要性会随之递减。但当人们对功用的分配进行衡量时，就会完全忘记这一点，以同样的思维对待功用和事物，将功利主义与财富主义相混淆。因此，数不清的哲学家对无辜可怜的功利主义者提出指责，宣告他们犯有危害人类罪。

（罗尔斯的反驳在“理论上”成立吗？***他从“最初动机”角度提出的观点合理吗？**）

## 公正与更大范围的利益

功利主义真的不公平吗？让我们回想一下。

功利主义是否要求我们将自己变成幸福之源？是否要求我们为了更大范围的利益而甘做奴隶？没有。因为道德英雄主义并不是我们的目标，这种要求对于活生生的人来说根本不现实。事实上，功利主义只要求我们尽量提高自己的道德水平，对社交圈外层的人们多些关心。功利主义并不要求我们成为道德完人，它只要求我们面对自身道德局限，然后尽己所能克服这种局限。科学也能够告诉我们，人类的责任感多么善变，多么不理性。

功利主义是否支持在刑事审判中颠倒黑白，让无辜者受罚，有罪者得奖？在想象的世界里，答案是肯定的；但在现实世界里，这些想法都是极其糟糕的，明智的幸福最大化支持者绝不会对任何一种情况表示赞成。同先前一样，这种协调也必然伴随着改革。人类对惩罚的偏爱有用，但并非绝对可靠。在奶昔随处可见的世界里，对脂肪和糖分的偏爱会让我们肥胖。同样的，对报复性惩罚的偏爱可能会造就一个刑事审判系统，满足我们对惩罚的偏爱，但这个系统却不利于整个社会的健康运转。

功利主义是否支持奴隶制或其他形式的压迫行为？在现实世界中并非如此。现实世界中，压迫所带来的幸福感十分有限，但被压迫者承受的痛苦却难以计量。人们之所以认为功利主义会造成社会不公，是因为他们的思维中存在财富主义谬误，将财富最大化与幸福感最大化混为一谈。“理论

上”讲，压迫他人确实有可能将幸福感最大化，但在现实世界中，考虑到人类的本性，压迫并不能让整个世界变得更加幸福。

因此，在现实世界中，幸福与公正之间并不存在本质的冲突。但通过理解正义感的认知原理以及工作机理，我们依然可以使自己的正义感更加完善。

# 第五部分 道德出路

# 第11章 深度实用主义

终于到了总结的时间，是时候把前文从生物学、心理学以及抽象的哲学角度所讲的内容进行整理，转化成某些实用的东西了。基于现有知识，我们应当如何看待引起分歧的问题？对那位拒绝购买医疗保险的蠢人，我们应当如何看待？医疗保险是一种权利吗？（蠢人也拥有的权利？）或者说医疗保险不过是一种帮人省钱的产品？10%的美国人控制了全国70%的财富，这样公平吗？或者说在这个充满机遇的国度，事情本该如此？全球变暖的威胁真实存在吗？抑或这只是空想社会改良家设下的骗局？如果真如专家所言，这种威胁真实存在，那么该由谁出资阻止情况恶化？出资金额多少才合适？伊朗是否有权利发展核技术？以色列是否有权利阻止他们？国际特赦组织将死刑称为“对人权的基本践踏”是否合理？首席法官亚历克斯·科津斯基（美国联邦第九巡回上诉法院）认为谋杀犯是“自己放弃了生命的权利”，这种观点正确吗？同性恋权利是逐渐成形的公民权利，还是上帝面前的丑陋行径？如果晚期病人想要结束自己的生命，医生应当提供帮助吗？抑或我们应当相信美国医学会，将医生协助自杀看作“背离了医者救死扶伤的根本职责”？

我们在第6章引入了一个绝妙的想法：作为新草地上的现代牧民，我们应当搁置各自的意识形态，共同选择最有效的方式生活。我们也已发现，这剂药方听上去十分合理，却并不好用。但基于之前的理解以及接下来的解释，这个想法其实非常强大，富有挑战。要想认真对待这个想法，就需要从根本上改变我们思考道德问题的方式。

## 两个罗盘

我将这种哲学思想称为“深度实用主义”，它给人的感觉十分平淡乏味，因为我们认为自己正在实践这种哲学，坚信自己的追求是出于最好的意愿。但毫无疑问，不可能所有人都是正确的。为了让这种哲学更具说服力，我们需要对“最好的”进行具体说明，我们需要一种共同的道德标准，也就是所谓的元道德。再次说明，元道德的作用是帮助我们做出艰难的道德抉择，在相互矛盾的部落价值观之间做出取舍。这个目标能够通过制定原则实现吗？

众所周知的“相对主义者”认为这是不可能的，因为不同的部落拥有各自不同的价值观，这个理由就已经足够。在根本的形而上学层面，相对主义者也许是正确的，道德问题也许根本没有绝对正确的答案，但即便果真如此，这样的认知对我们也没有任何帮助。法律必须立下一些规矩，我们必须做出选择，除非我们能接受掷硬币得出的结果，或者认为这样的结果就

是正确的，否则我们必须基于一定原因进行选择。不论道德标准怎样，我们的选择总要有个依据。

继续深入下去，有两种整体策略可供选择。第一种策略是寄希望于某种独立的道德权威，比如上帝、推理或者科学。正如第7章所述，这些道德权威都无法直截了当地提出道德真理，以解决我们面临的争端。因此，我们又被扔回了“泥潭”，重新面对纠缠不清的各种价值观和信仰，它们让我们团结一心，但同时又让我们分崩离析。

第二种策略是深度实用主义的策略，即从共同的价值观中寻找共同点，而不是依赖某个独立的道德权威。（上帝、推理或科学认为：“生命权超越了选择权。”）我们的目标是建立一种通用货币，用于衡量相互矛盾的价值观。功利主义基于人类体验，建立了一种通用货币，这正是该思想的精妙之处。如前所述，不论我们会因怎样的事情触动神经，体验都是重要的，既包括自己的体验，也包括他人的体验。人人都想得到幸福，没人愿意忍受痛苦。经过思考便不难发现，我们所看重的所有事的本质几乎都是对幸福和痛苦的关注。基于个人价值这个核心思想，我们可以对其公平地进行衡量，加入黄金法则的精髓，将其变成一种道德价值观：你的幸福和痛苦与其他任何人的幸福和痛苦不分贵贱，同等重要。最后，我们可以利用大脑前额叶皮层的结果优化机制对这种道德价值进行审视，将其扩大为道德体系。由此产生的道德哲学也许没人喜欢，但所有人都能够“理解”，这就是所有部落的所有成员都能够掌握的第二种道德语言。每个部落都拥有各自的道德直觉和自动模式设置，这也是部落之间产生争斗的原因。但幸运的是，我们都拥有灵活的手动模式。经过思考，我们便能使用手动模式思维在“大脑”层面达成一致，暂时搁置“心灵”之间难以调和的不同。这就是深度实用主义的精髓：依据事实而非想象寻找共同的基础。

实用主义这个词对我们并不陌生，有人赞赏，有人怀疑。赞赏实用主义者的人认为他们能够达成“结果”，能够实践“可能性的艺术”，还能够联结“我们”与“他们”之间的鸿沟。也有人担心，实用主义者满怀热情向前迈步时是否会迷失方向。假设两个孩子争夺一块蛋糕，一个孩子想把蛋糕从中间分开，但另一个孩子想要独吞蛋糕，这时出现了一个“实用主义者”，永远扮演促成和解的角色，他说：“好啦，孩子们。现在我们理智地想个办法。你拿走3/4，你拿走1/4。”不假思索地妥协并不是美德，有些折中是不好的，有些寸步不让却是好的。在寻求和解的过程中，如果我们将寸步不让的道德情感放到一边，那么剩余能为我们指明方向的是什么呢？我们的道德罗盘又在哪里？

深度实用主义一词中，“深度”十分关键。现代牧民无法通过简单的一句“我

们都理智些，寻求和解吧”来解决问题。实用主义者需要一种清晰连贯的道德哲学，需要在直觉不再可信时找到第二个道德罗盘，为我们指明方向。我花了大量篇幅对功利主义进行阐释、澄清、辩护，便是这个原因。我理解并非所有人都认为将功利主义付诸实践的时机已经成熟，但如果我们已经承认，所有部落的情感不可能全部正确，也渴望通过有序的方式解决分歧，那么我们必然需要某种“主义”，某种清晰的道德标准，以便在情感罗盘失效时为我们指引方向。

我们对功利主义并非一见钟情，正如第6章所述，功利主义很容易遭到误解。这种思想无关“功用”，并非要我们看重实用功能，看轻为生命赋予意义的事件；也不是浅薄地追求我们“最喜爱的事情”；这种思想并非自私自利，并非享乐主义，并非盲目的乌托邦空想。功利主义并不需要特异功能或者高科技手段来精确衡量幸福感，也不要求我们频繁进行“计算”。相反，功利主义有充足的理由反对这些天真的、伪功利主义的行为。真正的功利主义与讽刺者眼中的功利主义是全然不同的。如果能够正确理解功利主义，并将其恰当运用，我们便得到了深度实用主义，这是我们的第二道德罗盘，是新草地上生活的最好向导。本章将会讨论做一名深度实用主义者究竟意味着什么，还会提出一些看似不错的理论与之进行比较。

## 何时启用傻瓜模式？（我与我们）对阵（我们与他们）

前文中提到了两件明显矛盾的事情。一方面，我说应当将直觉放在一边，转而使用手动模式，依靠功利主义道德罗盘辨别方向。（把这些机械方面的比喻混在一起，我表示十分抱歉。）另一方面，我说深度实用主义者不应频繁进行功利主义计算。哪种说法是正确的？

这取决于我们所面临的是哪一种问题。整本书中含有三个主要比喻，这里涉及了其中的两个。第一个比喻是新草地上的寓言，阐释了常识道德悲剧的发生。第二个比喻是照相机的比喻，分别揭示了人类直觉反应（自动模式）与清晰推理（手动模式）的优点与缺点。为了回答“何时启用傻瓜模式”这个问题，我们需要将前两个主要比喻结合起来。（第三个比喻是通用货币的比喻，稍后还会提到。）

第一部分谈到，我们面临两个本质不同的道德问题。一个问题是“我”与“我们”之间的对立。如前所述，这是公地悲剧所反映的问题，也是合作的基本问题。道德思维会首先借助情感解决这个问题，同理心、爱、友谊、感激、荣誉、内疚、羞愧、忠诚、谦逊、敬畏以及窘迫等情感促使我们（有时）将他人的利益置于优先的地位。同样，愤怒与厌恶的情感促使我们回避甚至惩罚那些以自我为中心、忽视“我们”的人。如果没有自动模式，撒

谎、行骗、偷盗及杀戮等行为可能会比现在更多，“我们”的生存模式也不可能成功。

另一个更复杂的道德问题是“我们”与“他们”之间的对立，涉及我们的利益或价值观与他们的利益或价值观之间的对立，抑或是双方利益与价值观的同时对立。这是现代的道德悲剧——常识道德悲剧，也是新草地上争斗的起源。在新草地上，大家迥然各异的情感和信仰使和谐相处变得十分困难。首先，部族主义使我们以死硬的态度支持“我们”，忽视“他们”。其次，不同部落的合作方式不同：有些部落提倡集体主义，有些则更倾向于个人主义；有些部落以强硬的态度回应威胁，有些部落则强调和谐。再次，不同部落的“专有名词”不同，他们的领袖、文本、制度不同，为道德权威授予的行为也各不相同。最后，所有的这些不同最终导致人们对真实和公平的认识都带有各自的偏见。

第二个主要的比喻是照相机的比喻，阐释了道德思维的两种模式。人类大脑包含一种自动模式，即情感的本能反应，这种模式高效但相对死板；还有一种手动模式，是一种清晰的实践推理能力，这种模式十分灵活但相对低效。

我们面临两类道德问题，也拥有两种道德思维，现在可以对之前的问题进行回答了：要想明智运用道德思维，关键在于选取恰当的思维方式解决相应的问题。道德情感，即自动模式，总体来讲比较擅长抑制自私，避免公地悲剧。从生物学角度和文化角度来看，这也是自动模式最初的进化目的。因此，面对“我”与“我们”或“我”与“你”之间的矛盾时，应当信任道德本能反应，信任良心：不能撒谎或偷盗，即使手动模式为此找到合理依据也不行；不要偷税，也不要偷情；不要从办公室的钱箱里“借”钱；不要恶意中伤竞争对手；不要占用残障停车位；不要酒后驾车。如果有人做了这样的事，一定记得对他们表达你的蔑视。遇到“我”与“我们”对立的问题时，请信任自动模式。（道德的自动模式，而不是贪婪的自动模式！）

然而……面对常识道德悲剧，面对“我们”与“他们”之间的对立，便需要停止对本能反应的依赖，转向手动模式。可我们如何才能知道自己身处哪种情况呢？这个问题的答案极其简单：根据争议判断。如果某人的行为直接违反了道德规则，比如诈骗或谋杀，这属于道德问题，但其中不存在道德争议。法院外并没有抗议者为伯纳德·麦道夫欺骗投资人的“权利”抗争。这就是“他”与“我们”之间的对立，这种情况下，判断对错的直觉基本上值得信赖。

但如果争议出现，部落之间的意见向左，那么你就能知道，我们已经身处新草地中，面临“我们”与“他们”之间的问题，应当切换到手动模式了。为什么呢？部落之间出现不同意见，往往是因为不同部落在各自的自动模式

下持有的观点不同，因为不同部落的情感道德罗盘指向了相反的方向。这种情况下，常识不再能帮我们渡过难关，因为我们所谓的常识并不像想象中那样能够为众人所认可。

巧合的是，这里提到的决策策略（直觉发生冲突时，切换到手动模式）也是大脑在其他环境中使用过的策略。照相机的比喻为我们留下了一个关于大脑决策机理的谜题：摄影时，摄影师决定何时使用傻瓜模式，何时使用手动模式。但在人类的决策过程中，谁充当了摄影师的角色？如何决定依赖哪种模式做出决定？看上去我们似乎要返回看不到尽头的来路了。做出决定之前，我们不是先要决定以何种方式做决定吗？在我们决定以何种方式做决定之前，我们不是要……？

马修·波特威尼克（Matthew botvinick）、乔纳森·科恩（Jonathan Cohen）和同事们共同进行了一项开创性研究，揭示了大脑从该循环中脱身的方式。*你可能还记得第4章提到的斯特鲁普辨色任务，在试验中，不管一个词本身的含义是什么，参试者都需要对单词的字体颜色进行辨认。比如，看到用蓝颜色写出的“红色”这个词语时，你的任务是说出“蓝色”。这是很难的，因为与辨认颜色相比，阅读是更加无意识的反应。想要迅速准确地完成这个任务，人们便需要控制自己的认知，对手动模式进行控制。那么手动模式怎么知道自己该在何时介入呢？人们是不是每次选择思维方式前都要先问自己，“这种情况复杂吗？”

波特威尼克和科恩认为，大脑通过前扣带皮层（aCC）所控制的冲突监控机制来解决这个问题。一旦大脑中出现相互矛盾的反应，前扣带皮层便会兴奋起来。例如，看到用蓝色写的“红色”这个词时，一部分神经元开始兴奋，怂恿你说“红色”，而另一部分神经元则怂恿你说“蓝色”。根据冲突监控理论，前扣带皮层察觉到大脑中有两种相互冲突的行为，便向负责控制手动模式的背外侧前额叶皮层发出唤醒信号，该区域便像高级法院一样，出面调解争端。我、科恩与其他同事进行的研究结果也符合这一说法。我们在研究中提供了难以抉择的道德困境，其本身包含矛盾情感，我们发现，大脑考虑这类问题时，前扣带皮层和背外侧前额叶皮层都会兴奋起来。

斯特鲁普任务和道德困境中，矛盾与冲突都仅限在一个人的大脑内部。但牧民之间的争议是不同大脑之间的冲突。我的建议是：面对冲突时，请切换到手动模式。大脑面临脑内争议时，会自动采用这一策略，而面临不同大脑间的争议时，会有意采取这一策略。

## 源自深度

那么我们便假设，牧民们意见不一致时，我们需要停下来思考。很难，这真的很难，这是一个绝妙的想法，也是一个巨大的危险。考虑造成分歧的道德问题时，人们的第一反应就是寻找各种原因，证明“我们”是正确的，“他们”是错误的。

请再次回想前文中关于对待死刑态度的试验，关于死刑的威慑作用是否有效这一问题，试验人员向死刑的支持者和反对者提供了正反两方面的证据。尽管证据中包含了正反两方面的观点，但人们的观点不但没有变得中庸，反而更加极端了。人们只关注符合自己观点的证据，对其他的证据则不予理会。同样的，在气候变化的问题上，有很多能够自己阅读科学材料并进行“计算”的美国人，他们并不是气候变化方面的专家，只是喜欢使用自动模式的普通人，但他们的观点往往非常极端。当我们自己对各种证据进行衡量时，我们的偏见也会无意识地悄然而至。心烦意乱的谈判者将赌注压在毫无偏见的第三方上时，往往会亏损一大笔钱财。

了解了这一点之后，你可能会认为，有理有据的手动道德模式不可能存在，努力考虑这些分歧可能会让事情变得更糟。也许吧。或者也有可能，如果我们能够以正确的方式利用手动模式，这种思维便会让我们更加团结。现实世界中，大多数有争议的道德问题都非常复杂，比如全球变暖、医疗改革等。但在这些问题上，不是专家的人们往往都持有非常坚决的观点。在理想的世界里，我们会将自己变为专家，依据广博的知识做出判断。鉴于这种情况永远不可能发生，我们第二理想的观点就是模仿苏格拉底的智慧：认识到自己的无知，便会更加智慧。

心理学家弗兰克·凯尔（Frank Keil）和同事一同证明了“理解深度的幻觉”的存在。简而言之，人们往往认为自己已经理解了事情的运行机理，但事实上他们并没有理解。例如，多数人认为他们理解拉链或抽水马桶的工作原理，但一旦真的要他们试着解释，人们往往会失败得很惨。但最关键的是，一旦人们试着解释某件事的工作原理但未成功，他们便会意识到自己的失败，然后对自己的认知水平进行重新估计。

在一系列设计精妙的试验中，菲利普·费恩巴赫（Philip Fernbach）、托德·罗杰斯（Todd rogers）、克雷格·福克斯（Craig Fox）以及史蒂文·斯洛曼（steven sloman）将这个观点应用到了政治领域。他们让一些美国公民对6个富有争议的政策提议进行考量，包括医疗系统中的统一支付制度和碳排放交易体系。一次试验中，试验人员要求参试者描述自己对这些政策的看法，并表明自己对该政策的理解程度。随后，他们要求参试者对这些政策的原理进行详细解释。最后，参试者需要再次描述自己对政策的看法，评估自己对该政策的理解程度。结果发现，人们被迫对这些政策的运行机理进行解释之后，会降低自己对政策理解程度的估计，对该政策的态度也会更加缓和。在这个试验的对照组中，参试者无须解释政策的原理，而是

要为自己的观点提供论据。对于大多数人来说，提供论据并不会对其坚决的态度造成影响。

这些研究完美诠释的是，正确的手动模式思维能够使我们更加团结。让人们以清晰的论据证明其观点并不能够使人们更加理智，甚至可能产生相反的效果。但迫使人们面对事实，承认自己不了解关键事实，却能让人们的态度趋于缓和。研究者认为，这些结论为公共辩论的方式提供了另外一种可能性。我们可以不再简单地向政治家和权威者询问他们为何支持某些政策，而是让他们解释自己所支持政策的原理。适用于“与媒体见面”节目的，也同样适用“与家属见面”。如果你有一位固执己见、油光满面的叔叔，像我们所熟知的那样坚持认为国家医疗保险政策是历史的进步，或者认为这项政策意味着文明的终结，你也许不必与他直接对抗，就能够改变他的想法。你可以对他说：“吉姆，你的观点很有意思。那么国家医疗保险体系究竟是如何运作的呢？”

## 心灵的秘密玩笑：合理化思维与大脑的双加工机制

20世纪70年代早期，唐纳德·达顿（Donald Dutton）与阿瑟·艾伦（arthur aron）在试验中安排了一位美丽的女子在不列颠哥伦比亚省一所公园的两座桥上分别与人搭讪。一座桥横跨深谷，是摇摇晃晃的吊桥，看上去让人心惊胆战。另一座桥则比较矮，是稳固的木桥。这位美丽的采访者会拦住男性被访者，询问他们逛公园的感受，并且留下自己的电话，如果男性被访者想要更多地了解这项研究，可以打电话询问——眨眼暗示。结果发现，吊桥上遇到的男子中，大多数人给她打来电话，也有很多人想要约她出去。为什么呢？达顿与艾伦推测，这是因为吊桥上的男子将自己心跳加速、掌心出汗等表现当成了心动的感觉。结论是：人类不知道自己的感觉因何而起时，往往会编造一个貌似有理的故事，让自己相信。

这种现象并不是孤立的。在另外一个经典试验中，理查德·尼斯贝特（richard Nisbett）与蒂莫西·威尔逊（Timothy Wilson）让参试者从一排连裤袜中选出一双。随后，当人们被问及为何选择这双而不是另外一双时，他们都给出了十分合理的答案，提到了所选袜子的很多特点，包括精细的织工、是否透明、弹性如何等方面。但事实上，人们做出的选择与这些特点完全无关，因为所有备选的袜子都是完全相同的。人们不过是偏爱放在右手边的事物罢了。这两位科学家还做了另一个类似的试验。他们先向人们展示很多组词语，其中一组是“海洋—月亮”。随后，他们要求参试者从不同品牌的洗涤剂中进行选择。之前看过“海洋—月亮”这组词语的人们选择“汰渍”的概率比其他组高出一倍。但当参试者解释自己的选择时，他们给出的理由是“汰渍是最有名的洗涤剂”、“我的母亲就用汰渍”，或者是“我喜欢汰渍的包装”。

人们倾向于为自己做出的事情编造理由，神经病患者往往无法解释自己的行为，他们的应对方式有力地说明了这一点，例如，Korsakoff遗忘综合征患者常常试图编造故事，掩盖自己的记忆缺陷。他们在编故事时往往自信满满，完全意识不到自己在胡编乱造。神经病学家将此称为“虚构症”。例如，一次研究中，一位遗忘症患者坐在空调旁边，试验人员问他是否知道自己身在何处。他说自己在一家空调制造厂里。试验人员指出他穿的还是睡衣，他回答道：“我的睡衣就放在车里，一会儿我就换上工装。”为了避免癫痫恶化，医生有时会人为地通过手术切断大脑两个半球之间的联系，这种患者被称为“裂脑患者”。在裂脑患者身上也会出现与遗忘症患者类似的情况。通常情况下，大脑的两个半球都能够知道另外一半在想什么，但联系切断后，这种内部信息便会缺失。一次研究中，试验人员向病人大脑的右半球展示了一张雪景图片，并要求病人选择一张与之匹配的图片。由大脑右半球控制的左手选择了一张铲子的图片。与此同时，试验人员向控制语言能力的大脑左半球展示了一张画有鸡爪的图片。之后，试验人员口头询问病人为何会用左手选择铲子的图片。病人大脑的左半球只看到了鸡爪，但没有看到雪景，于是病人答道：“我看到了一只爪子，然后选择了铲子，因为清理鸡棚需要铲子。”

“虚构症”是很奇怪的病症，但认知神经学家从中得出的结论则更加奇怪：大脑损伤并不是导致或激发虚构症的原因。毕竟，大脑的损伤不可能为其赋予新的能力或动机。事实上，科学家们得出的结论是：我们所有人都是虚构症患者，大脑健全的人甚至更加擅长于此。我们总在不断地解释自己的行为，将我们所做的事以及这样做的原因加以修饰，使之成为一个看似合理的故事。神经学上的虚构症患者与我们其他人之间最大的区别就是，患者由于神经缺陷，被迫基于更少的线索架构自己的故事。如果想要让健康的人暴露自己的虚构症行为，就必须通过对照试验，比如前文提到的桥梁试验和汰渍的试验。

在道德层面，与虚构症相对应的是合理化思维。虚构症患者感知到自己在做某件事，便会据此编出一个看似合理的故事，解释自己的行为及其原因。进行合理化思维的主体因为某个道德问题产生一种感觉，便会据此为这种感觉找到一个看似合理的理由。乔纳森·海特（Jonathan Haidt）认为，每个人都有完美的道德合理化思维。这是大脑思维双加工机制的必然结果。自动模式为我们提供情感上具有说服力的道德答案，随后，手动模式开始工作，为这些答案寻找合理的解释。这种做法与健忘症患者相似，都在为自己的行为寻找理由。下面一段话摘自伊曼努尔·康德的《论过度自慰》（*Concerning Wanton Self-Abuse*），阐释了康德为何认为自慰行为是对绝对命令的违背，违反了最高道德准则。

这种行为是对性别属性的非正常使用或滥用，违背了一个人对自己的责

任，是最严重的违背道德的行为。所有人一想到这一点，都会有所触动。……但要想理性地证明这种行为是对性别属性的非正常使用，是不可接受的行为，是对自身责任的违背，却并非易事。……如果一个人将自己仅仅作为满足动物本性的一种工具，那么他就放弃（抛弃）了自己的人格。这个事实无疑可以作为一项论据。

根据阿奎那的双重效应原则，手段伤害与连带作用伤害是不同的。类似的，康德认为如果某种行为将人用作了工具，那么这种行为就是错误的。康德将这个观点用于批判手淫的罪行：因为手淫者将自己作为工具，所以这种行为是错误的。

这是聪明的想法，但却有些滑稽。康德试图通过抽象原则证明自慰的不道德。反对这种压抑性欲习俗的人可以对康德这种热切的努力尽情嘲笑。19世纪德国哲学家弗里德里希·尼采（Friedrich Nietzsche）也认为康德这种理性主义的道德观很滑稽：

康德的玩笑——康德想要通过令平民感到震惊的方式，证明平民的正确性。这是他心中的秘密玩笑。他写文章反驳学者的观点，支持多数人的偏见，却是为学者而写，而非为人民而写。

也就是说，康德的自动模式设置与他身边同部落的人完全相同。但与他们不同的是，康德认为自己需要为“多数人的偏见”提供深奥的理由。康德还建立了一套完善的理论体系，证明白种人的优越和黑种人的低劣。他认为黑人是“天生的奴隶”。

合理化思维是道德发展道路上最大的敌人，因此也是深度实用主义最大的敌人。*如果道德部落之间相互争斗是因为内部成员的直觉各不相同，那么通过手动模式为情感寻找合理化理由将完全无益于解决问题。我们需要使用手动模式，但我们需要明智地使用。我们对此已经有了一些头绪（基于“为什么”，我们解释了“怎么办”），但我们可以做得更好。我们可以学着辨认合理化思维，制定基本原则，让自己更加不易被愚弄，也让彼此之间更加不易受骗。

## “正面我赢，反面你输”：合理化思维中的权利思维

作为深度实用主义者，我们一心扑在艰难的实证工作上，想要寻找在现实世界中最为有效的体制。但部落的效忠派依靠自己永远正确的直觉，必然会反对我们的立场。反对死刑的人们乐于引用一切可能的证据，向你证明死刑无助于降低犯罪率。而死刑的支持者则会干劲儿十足地做出相反的行为。对于部落的效忠派来说，这些实用主义或功利主义的论据都不过是绣花枕头。如果死刑真的像国际特赦组织所定义的那样，是“对人权的基本

践踏”，那么在政策层面上的争论无异于“正面我赢，反面你输”的规则。如果事实证明死刑不好，那么国际特赦组织便会欢呼雀跃。但如果事实相反，死刑在“理论层面”依然是错误的。毫无疑问，死刑的支持者也会采取同样的策略。当经验主义的对抗上升到言语的冲撞时，他们便会坚持认为，死刑是社会的道德权利，蒙冤已久。

因此，“权利”是学者们最好的通行证，也是一张王牌，永远能拿出不相干的证据。不论你和部落中的同伴产生了怎样的感觉，你们都能够找到相应的权利，与这种感觉相对应。如果你感觉堕胎是错误的，便可以诉诸“生命权”。如果你感觉禁止堕胎是错误的，便可以诉诸“选择权”。如果你是伊朗人，可以诉诸“核的权利”；如果你是以色列人，便可以诉诸“正当防卫的权利”。将“权利”作为理由无疑是高明的，我们可以毫不费力地为一切直觉找到正当的理由。

权利和义务就像是镜像的关系。在现代道德辩论中，两者是完美的言语武器。如前文所述，大脑的自动模式发出道德指令，告诉我们某些事应当做，某些事不能做。这些感觉与权利和义务的概念基本能够完美契合。如果我们感到某件事是错误的，便可以说这件事侵犯了人类的“权利”，从而将这种感觉表达出来。同样的，如果我们感到某件事是必须要做的，便可以援引相应的“义务”表达这种感觉。将人推下天桥让我们感觉非常糟糕，所以不论这种行为是否会拯救5条生命，我们都会说这是对人权的粗暴践踏。然而扳动开关给我们的感觉并没有那么糟糕，因此我们说这种行为并未侵犯受害者的权利，或者说受害者的权利不及被救5人的权利重要。*同样的，我们有义务挽救身旁的落水儿童，但远方“以数字代表的”儿童并没有让我们感到如此触动，因此我们便没有挽救他们的义务。权利和义务是由情感来支配的。

权利和义务的表达能够巧妙地表达道德情感，具体体现在两个方面。第一，当直觉提出必须做和不能做的事情时，这些命令是不容置疑的，体现了自动模式的刻板。命令我们不得将人推下天桥的情感并不在意有多少生命危在旦夕，0人也好，5人也好，甚至100万人，结果也不会改变。我们可以忽视这种情感，但就其自身而言，并没有商榷的余地。一位实验心理学家将其称为“认知壁垒”。这种不容置疑的态度深植于权利和义务的概念当中。权利和义务可以被忽略，但这绝对不仅仅是打破认知平衡这么简单。权利和义务是绝对的——但有些时候则不是。

第二，论战中的道德家喜欢援引权利和义务，通过对客观事实的理解，表达人们的主观感受。我们对这种方式钟爱有加，因为我们总认为主观感受来源于对显而易见事实的感知，尽管有些时候却并非如此。以异性产生的吸引力为例，如果你认为某人十分性感，其实就是你的内心将性感的光环投射到了外在的对象身上。虽然我们不会这样认为，但事实就是如此。我

们当中有些人会觉得另外一些人十分性感，但大部分人都不会认为一只狒狒十分性感。毫无疑问，狒狒也只会对狒狒感兴趣，而不会对人类感兴趣。同时，部落内部的分歧也提醒我们，性感与否只在于观察者的判断。*但不管怎样，当一个人被异性吸引时，他/她可完全不会这样想。我们并不会将性感看作观察者内心的主观投射。在我们看来，性感像身高和体重一样，是显而易见的事实。因此，我们可能会认为某人十分“性感”，但不会说这个人“激起了我们的性欲”。同样，权利和义务的说法跳过对客观事实的考量，将主观感受作为显而易见的客观事实进行描述。当你声称某人拥有某项权利时，你的态度就像是对这个人拥有的一切进行客观描述，就像在说她有十根手指一样。

如果我是正确的，大脑的手动模式试图将抽象的情感转化为更加具体的事物，以便进行理解和控制。权利和义务便是这种尝试中诞生的产物。手动模式存在的首要意义是为了解决外部世界的事件：行为、事件，以及联系两者的因果关系。因此，手动模式的本体是具体的“名词”和“动词”。但人类还有很多不知从何而起的神秘情感；与合理的方式反其道行之的抗议行为；或是强制自己完成可有可无的行为等，这些都是自动模式的调控结果，可是手动模式应当如何对其做出合理解释呢？结论：这些情感作为对外界事物感知的结果而出现，尚且没有统一的定论。模糊的不应做某事的感觉被人理解为“权利”。“权利”是一种抽象但却真实存在的事物，可以与“获得、失去、放弃、转让、扩大、限制、比较、取消、威胁、交易、违反、维护”等动词搭配。我们将自己的道德情感概念化，转变为对权利和义务的理解，使通常用于具体事物的认知系统能够派上用场，使我们能够更加清晰地思考这些概念。

因此，出于以上所有原因，权利和义务被现代道德家当作选择的武器，我们得以将自己的情感变成不容置疑的事实表达出来。通过对权利的援引，我们逃脱了责任，不必再为自己想要的东西苦苦寻找真实可信、直截了当的理由。只要我们放任自己打出权利这张王牌，证据便是无关紧要的，因为游戏规则是“正面我赢，反面你输”。

至此，也许你会认为我对权利的看法过于极端：上述观点反对一切对权利的援引吗？还是仅仅对凭空胡乱援引权利概念的行为提出反对意见？我们当然可以用权利的概念将我们的直觉变得合理，但我们也可以试着通过功利主义思想达到同样的目的：不论心中想要什么，我们都说这是为了更大范围的利益。这两种方式有什么区别吗？

如前所述，与对权利的援引不同，是否能够促进更大范围利益的考量最终建立在证据的基础上。一项给定的政策是否能够增进幸福感是需要通过事实检验的。我们可以声称国家医疗保险制度将会促进或毁掉美国医疗体系，但倘若我们想要底气十足地提出这样的观点，能够拿出一些证据是最

好的。首先，我们最好能够理解国家医疗体系的工作机制（参考前文内容）。然后，为了寻找证据，我们必须弄懂各种医疗体系之间的区别，了解不同国家医疗体系的实施状况：哪国人的寿命最长？哪个国家提供的终身护理质量最高？哪个国家的公民对其医疗体系的整体满意度最高？当然，这些都是政策专家们需要回答的问题。这些问题不仅适用于医疗体系，也适用于一切重大社会问题：某国取消死刑惩罚时，谋杀率升高了吗？某国在更大范围内对财富进行了重新分配，这是对懒惰的鼓励吗？这些国家的公民整体幸福感降低了吗？要想找出使社会整体上更加幸福的因素十分困难，而且很容易得出带有偏见的结论。但最终，我们前进十步，倒退九步，便能够拿出证据，回答这些问题。

然而，关于权利的问题却无法得到类似的回答。第7章中提到过，至今我们无法直截了当地证明何人拥有何种权利。如果有一天，哲学家能够建立一套正确的权利理论，那么这里我所说的一切便都可以被丢到窗外去了。但至少现在，诉诸权利并无益于解决问题，是没有出路的。相反，这种做法是对自己的欺骗，假装问题已经在某个抽象的领域得到了解决，而这个抽象的领域只有你和你的族人能够进入。

长期以来，你都对权利的观念坚信不疑。因此现在你对我的观点也许依然半信半疑。你也同意，大部分关于权利的言论都是空洞的借口，但权利的概念似乎反映了一些非常重要的本质，而这些本质在功利主义的收支账目表上却无法得到体现。将女童卖作娼妓与权利无关吗？对勇于表达自己信仰的人们施以酷刑与权利无关吗？这些事件难道不是对人类权利的践踏吗？我们的道德罗盘去哪儿了？

我有个好消息要告诉你。作为深度实用主义者，我们能够认识到权利思维在道德生活中的重要性，过去如此，未来也将继续如此。为权利而争论也许毫无意义，但有时争论本身便是毫无意义的。有时候，我们需要的并非争论，而是武器。*这便是我们为权利辩护的最佳时机。

## 权利：矛与盾

法学教授艾伦·德肖维茨（alan Dershowitz）曾经为一些学生讲述了这样一个故事。一位犹太大屠杀的否认者要求与德肖维茨进行公开辩论，遭到了德肖维茨的拒绝。于是这个人不断给德肖维茨写信，表达自己的愤怒，质疑德肖维茨的学术品格。“你自诩为自由演讲之王，却想要让我保持沉默！你为何反对与我进行公开交流？你知道我会在辩论中胜出，所以根本不敢与我辩论！”最终，德肖维茨同意了这个人的请求，他说：“我同意与你辩论，但有个条件：我们的辩论必须分为三部曲。首先辩论地球是否是平的，然后辩论圣诞老人是否存在。之后，我们再来辩论犹太大屠杀是否

曾经发生。”至此，这位蓄势待发的辩手拒绝了德肖维茨的提议。

德肖维茨的答复是巧妙的，为我们揭示了一条有价值的实用主义观点：道德辩论不仅仅是对真理的探求。是否迎战对手，以及以何种策略出战，这些都是包含了成本效益分析的实用主义决策，与其他的决定没什么不同。在德肖维茨的例子中，需要比较的一方面是进行公开论战的好处，另一方面则是在一名心怀恶意的怪人身上所浪费的时间和精力。*有些问题毫无辩论价值，比如这个例子中的历史事实，但有些价值观也同样不值得争辩。

简单的网络调查显示，现今依然有人认为黑人理应受奴役；有些女人理应被强奸；希特勒没能灭绝犹太人实在遗憾。同样的，我们没有必要与这些人进行争辩。现代牧民已然达成共识，奴隶制、强奸和种族灭绝政策是绝对不能被容忍的。也许我们有不同的方式对此进行解释。有些人援引上帝的意志；有些人援引人权作为依据；还有些人与我一样，认为这些行为会造成无谓的、难以想象的痛苦。还有很多人，也许是绝大多数人，只是简单地表示反对，并将这种反对作为道德常识，并未考虑过具体的解释理由。但不管怎样，我们都同意，对这些行为不能有丝毫姑息。也就是说，人类在道德判断层面确实达成了某些共识。共识不代表全民赞同。共识意味着有足够多的人支持这个观点，意味着我们可以据此制定实际的政治目标，也意味着这个问题已经得到了解决。

当我们谈及已经真正得以解决的道德问题时，援引权利概念是完全合理的。为什么呢？因为权利的语言能够恰如其分地表达出我们坚定的道德决心。对某些观点的坚定支持与对其他观点的果断拒绝都十分重要。*这并非因为我们对所有问题的判断都绝对正确，而是因为犹豫不决比判断错误的风险更大。我们希望自己的子女能够在理智与情感的层面同时明白，有些事情是绝对不能触及的。我们想要向人群中的极端主义分子——3K党成员、新纳粹分子、厌女主义者——发出明确的信号：这里不欢迎他们。

我在前文中提到，我之所以反对奴隶制，是因为奴隶制的成本大大高于其收益。听到这种说法，你是否会感到有些不舒服呢？其实这样说，我自己也有些不舒服。这种说法让人觉得如果有人能够提出恰当的论据，我对奴隶制的态度便有可能，只是有可能，发生改变。不过请放心，在这个问题上，我的态度是无可动摇的。如果我看到一封标题为“论奴隶制在某些情况下的合理性”的邮件，我会表示感谢，直接点击“删除”。如前所述，我依然相信反对奴隶制最直接的理由来自于功利主义思想，由边沁和密尔在多年前提出。但身处新千年的时代，我却非常乐意以“迂回”的态度对待奴隶制的问题。据我估计，倘若将奴隶制视为一个开放性问题，然后通过手边的证据加以解决，这样做的成本远远超出其收益。因此，作为一名深度实用主义者，我很乐意加入人群，一同呼喊：奴隶制践踏基本人权！

也许你会提出反对意见："但其实你并不是这么想的！"是的，你是对的。但对于深度实用主义者来说，以适当方式宣告的人权观点就像是婚礼上的誓言。当你对爱人说出"至死不渝"时，只要你是一名头脑清醒的成年人，只要大脑的手动模式能够正常工作，你的意思并不是说你在任何情况下都不会提出离婚，也不是说人为选择结束这段婚姻的概率是百分之零。你只是在表达一种情绪，表达一种坚定的承诺。倘若站在圣坛之上，你说出口的是："亲爱的，据我估计，我们两人相守一生的概率非常非常高。"这种表达情绪与承诺的方式简直糟糕透顶。同样的，如果我们说"根据估计，奴隶制显然无法将幸福感最大化"，这也不是表达反对奴隶制立场的恰当方式。如果有人问你，"你是否相信奴隶制践踏了基本的人权？"正确的回答是："我相信。"

作为深度实用主义者，当道德问题得以解决时，我们便可以援引权利概念。也就是说，对于权利的援引可以作为盾牌，保护我们的道德进步不受威胁。同样的，有时我们也可以将"权利"作为武器。在论证失效时，"权利"可以作为言语工具，促进道德的进步。例如，在美国民权运动的道德斗争中，有人从功利主义角度提出论据，证明黑人也应拥有投票权，拥有在餐厅与白人同桌用餐的权利。这些论据都非常有力。但同所有的功利主义论据一样，这些论据基于公正不倚的前提，基于黄金法则，基于所有人的幸福同等重要的认知。这些前提恰好正是民权运动的反对者们所排斥的观点。因此，公开种族歧视这一问题与税收高低、死刑问题以及医生协助自杀的问题不同。站在公平不倚的道德角度，这个问题完全没有争论的必要。吉姆·克劳法（种族隔离法案）简单地反映了一个种族对另一个种族的统治，*直到20世纪50年代，我们意识到，单凭道德推理显然无法解决这一问题。我们需要的是外力推动，需要借助来自第三方的情感承诺。因此，在这个重要的道德与政治斗争阶段，带有显著情感特征的权利概念是最合适的语言。也许这个问题并没有得到解决，但同时，双方也不再有理性辩论的空间了。

因此，深度实用主义者有时也可以自由援引权利概念，既包括法律权利也包括道德权利。但事实上，这种场合比我们想象中还要少。如果我们真的想要理性地将反对者说服，那么我们就要避免使用权利概念。这是因为我们无法直截了当地（不借助功利主义思想）辨明何种权利真实存在，不同权利之间的优先等级又是怎样的。但如果某个问题已经得以解决，或者反对者无法理性思考，这时的争辩毫无意义，那么我们就应当停止争辩，集结言语的力量，进一步巩固道德承诺。靠不住的概率估计在这个过程中并无助益，打动心灵的言语才有效果。

但请务必不要将这个观点作为依据，将我提到的一切关于"权利"的事物全部忽略。大多数的道德论战的起因并不像一个种族统治另一个种族这样简

单。事实上，几乎所有的道德论战中，双方都有真正的道德考量。*提倡个人主义的社会体系有其优越之处，因为它鼓励人们照顾自己。而提倡集体主义的社会体系也同样有其优越之处，因为每个人都能够得到所需的帮助。禁止终结人类胎儿生命有其合理之处，让人们自己面对艰难的生物伦理学抉择也有其合理之处。用发自内心的权利主张彼此攻击并非解决之道，尽管这种解决方式看上去十分诱人。我想再次强调，正确的解决方法应当是将自动模式搁置到一边，转而启用手动模式，以通用货币为媒介，试图达成共识。

## 案例分析：堕胎

关于堕胎的争论由来已久，也激烈异常。因此，我们可以将堕胎作为一个案例，用于检验深度实用主义思维。如果深度实用主义思维能够在这件事上提供帮助，那么在其他的事件上，这种思维应当也是有用的。（我必须强调，我并非采用这种思维方式的第一人。这一部分和下一部分的内容中，有很多观点源自彼得·辛格等其他学者。）

道德调节者认为，我们所有人都应当更加理性、更加灵活、更加大度。这意味着什么呢？倘若你认为堕胎等于谋杀，等于终结了一条无辜的生命，你是否应当“理性”一些，允许别人进行谋杀？如果你认为禁止堕胎侵犯了女性的基本权利，你是否应当“理性”一些，放弃女性的选择权利？简单地要求人们更加理性并无助于解决问题，因为所有人都认为自己已经足够理性。要想取得真正的进步，我们就必须将直觉反应放到一边，切换到手动模式。事实证明，不论是左翼还是右翼，几乎没人能够就堕胎问题取得道德共识，任何现有观点都经不起手动模式的推敲。

我们先从维护选择权的一方开始。我们知道，自由派倾向于将堕胎看作事关“权利”的问题，具体说来，是关乎女性权利的问题，但几乎没人认为女性有权放弃9个月大的胎儿。这是为什么呢？9个月大的胎儿依然在母体之中，难道女性无权控制自己的身体吗？来自南方保守区域的基督教基要主义老年议员难道有权对旧金山的年轻女性指手画脚，禁止她们选择堕胎吗？从某种角度来说，他们确实有这样的权利。

为了避免前后矛盾的观点，选择权的维护者必须解释，为何怀孕早期堕胎合乎道德，但怀孕晚期堕胎却恰恰相反。*3个月以内的胎儿与6~9个月大的胎儿都具有发育成人的潜力。因此，两者在道德层面上的区别与潜力无关。不论堕胎发生在怀孕的早期还是晚期，都有一个人的生命被终止了。*如果潜力不是关键因素，那么想必关键因素在于事实状态：在怀孕的早期和晚期，胚胎实际状态的差别。可能的原因包括很多。

在著名的罗伊诉韦德案（roe v. Wade）中，美国联邦最高法院针对怀孕早

期和晚期的胎儿，提出了最具影响力的区分因素：胎儿离开母体后能否存活。但这个因素真的是决定性因素吗？胎儿离开母体后能否存活，一方面取决于技术水平，另一方面也取决于胎儿本身。*今天，即使是22周的早产儿也能够存活。随着技术发展，这个时间节点也必然会随之发生变化。也许在我们有生之年的某天，苦苦求生的胎儿能够在怀孕早期便脱离母体，在人造子宫中进行发育。若果真如此，选择权的维护者是否会说，由于新技术的发展，怀孕3个月以内进行的堕胎变成了违反道德的行为？*那些已到孕晚期却依然无法在子宫外存活的胎儿呢？假设一个不足9个月的胎儿，因某些特殊的状况依然无法在子宫外存活。假如这种状况在临盆之前便能得到改观。那么由于这个胎儿（尚且）不能在子宫之外存活便将其扼杀，这种做法合乎道德吗？

能否在子宫外存活这个因素似乎只是关键因素的一个便捷的代表。那么真正关键的因素是什么呢？怀孕后期胎儿所独有的、赋予它们以生命权的那个特征究竟是什么呢？探寻的过程也许会很艰难。因为不论这个特征是什么，我们大多数人所食用的动物也几乎一定会具有这个特征。是感受疼痛的能力吗？猪也能感受到疼痛。（不论怎样，我们至少能够确定，成年猪与怀孕晚期胎儿一样，都能够感受疼痛。）同样，与人类的胎儿相比，至少猪更可能有意识；更可能产生强烈的自我意识；更可能拥有复杂的情感；更可能与其他个体产生有意义的联系。在道德层面上，怀孕早期与晚期的胎儿之间产生的显著区别，在成年猪和其他作为人类食物的动物身上也同样存在。

这并非意味着选择权的维护者已经走投无路，但可走的路确实不多了。有一种途径是证明孕晚期胎儿拥有某种特质（如：意识初步形成），因此在怀孕后期堕胎违背道德，同时，食用某些动物也是违背道德的。证明这一点也绝非易事。为了避免自相矛盾，成为一名道德上的素食主义者还远远不够。*激进的素食主义者才能够满足条件。很多素食主义者，包括出于道德动机的素食主义者虽然自己不吃肉，但依然保有对吃肉的“选择权”。他们不会把食肉的朋友看作谋杀犯，也不认为食肉应当为法律所禁止。（有些人真的这样认为，但多数素食主义者并非如此。）如果你认为孕晚期的胎儿已经初步形成意识（或者其他什么原因），因此孕晚期的女性不应拥有堕胎的选择权，那么在吃猪肉这件事上，你也不应拥有选择权。这种立场只是一种选择，但绝大多数的选择权维护者是不愿走出这一步的。

另外一种途径：你可以认为孕晚期的胎儿身上汇聚了奇妙的特征，这种奇妙的特征赋予它们以生命权。同猪一样，胎儿拥有初步意识（或者其他什么特征）；但与猪不同的是，胎儿属于人类。同孕早期胎儿一样，孕晚期胎儿也属于人类；但与其不同的是，孕晚期胎儿已经拥有了初步的意识。你可以说，两个因素中的任何一个都不足以将生命权赋予某个个体，但当

两者相结合时——“啪”的一声——我们就得到了拥有生命权的个体。关于这个理论，我们需要注意的第一点就是，它是一种极其特殊的情况。其次，尤为特殊的一点是，这个理论强调了人性本身的关键性。几乎没有自由主义者会宣称只有人类才能拥有生命权。例如，很多自由主义者认为，除人类以外的其他动物，比如黑猩猩，也拥有生命权。我们不能出于一己私利将其杀死。为了加深理解，我们可以考虑外星人的道德权利，它们与人类具有同样的思考和感知能力。例如《星际迷航：下一代》（*Star Trek： The Next Generation*）中可爱的迪安娜·特洛伊，我们当然不能因为她不是人类就将她杀死。*令星际迷航迷们感到忧伤的是，特洛伊并不是真实存在的。但我们依然可以用她作为例子，支持这个观点：让我们拥有权利的特质并不是人类的身份，而是人类和其他物种拥有或能够拥有的特征。

将“人类意识”作为关键性因素的观点让我们想起了另一个类似的观点：灵魂的拥有。稍后谈到生命权的拥护者所面临的窘境时，我们会对这一点详加论述。但首先，我们可以设想选择权的拥护者会如何利用灵魂的观点为自己辩护。我们可以假设人类是有灵魂的，而其他动物或者没有灵魂，或者拥有本质上区别于人类的灵魂，比如猪灵魂等。再假设拥有（成为）人类的灵魂便意味着明确的生命权。如果某位选择权的拥护者拥有这样的想法，他会认为，孕晚期的胎儿已经拥有了灵魂，而孕早期的胎儿则没有。但这种观点的问题在于，没有任何理由能够证实这一观点。孕早期的胎儿也能够活动身体，也是有生命的。如果赋予胎儿生命的不是灵魂，还能是什么呢？暂时的胎儿灵魂吗？不管怎样，即使我们能够证明灵魂的产生，也无法理直气壮地宣称，灵魂的产生必定是在怀胎三月以后发生的。如果我们认为人类拥有灵魂，而孕早期的胎儿也拥有灵魂，那么这个观点对选择权的拥护者来说就不是有力的论据了。

总之，选择权的拥护者很难找到前后一致的合理解释来维护自己的立场。这并非不可能完成，只是即使这样的解释真实存在，我们也需要依靠手动模式进行复杂的哲学诡辩，建立一种晦涩的理论。在流行道德话语中，简单地宣称“我相信女性应当拥有选择的权利”就足够了，无须过多地解释。但如果没有进一步的解释，这种诉诸“权利”的论证方式不过是虚张声势，是一种单薄的主张。所能达到的效果就是，人们知道，世间存在着一种清晰的主张，拥护选择权，拥护人类的生殖权利。

那么生命权的拥护者们有何想法呢？他们的立场是不是更好些？拥护生命权的其中一种观点重点考虑了由于堕胎而无法真正开始的那条生命。问题在于，这个观点中混入了太多主观的因素。堕胎确实剥夺了一个人的生命权，但避孕措施也是如此，而大多数支持生命权的反对堕胎者尚未打算将避孕措施看作违法行为（至少在美国，情况如此）。很多拥护生命权的人

确实是虔诚的天主教徒，对避孕持否定看法，但问题并没有就此而止。剥夺生命权的观点也同样适用于禁欲行为。如果一对夫妻决定不生孩子，或者决定少生孩子，那么他们也是在剥夺孩子的生命权。即使对无力养活更多孩子的夫妻来说，只要有人愿意收养他们的孩子，他们不生更多子女的行为也是对生命权的剥夺。除非你认为让尽可能多的婴儿快乐生活是一种道德要求，若非如此，你便不能以堕胎剥夺了人的生命权为论据，证明堕胎的错误性。

然而，这个思路并不能代表多数生命权拥护者的想法。多数拥护生命权的人想要在生命的可能性与正在形成的生命之间划出一条界线。他们认为，受孕是一个关键的时间点。（下文中“受孕”与“受精”两个词语均用来指代精子与卵细胞的结合。）人们常说“生命”从受孕一刻开始，但严格来讲，事实并非如此。毫无疑问，形成受精卵（胎儿发育伊始的单一细胞）之前，精子和卵细胞都是有生命的。所以说，生命并非始于受孕一刻，而是某人的生命起源于受孕之时。这样的说法正确吗？

话题又回到了灵魂的问题上。但在此之前，我们可以先试着以不那么抽象的方式解释这件事。你可能会认为受孕过程是特殊的，因为当精卵相遇的那一刻，一个人的身份便从此决定。“谁的生命受到了威胁？”这个问题的答案也就此产生。但事实上，回答这个问题并不需要等到精卵真正结合。在生育诊所中，受精这个过程便发生在身体之外。通常情况下，受精容器（多为有盖培养皿，而非人们通常以为的“试管”）中放有多个卵细胞和许多精子，其中只有一对会成为幸运儿，使得受精过程成功。生育诊所的医生也可以选出一个精子和一个卵细胞，让它们完成结合。在此之前，两个细胞被分别放在不同的容器中，当幸运的精子“整装待发”时，如果此去能够成功，那么未来的那个人究竟是谁，在此刻便已经决定，*准受精卵的基因在此刻也已经决定。但在这个时刻，拥有生命权的那个人是被分成两份，在两个不同的容器当中吗？如果精子已经就位，但那位女性选择了退出，这种行为算作谋杀吗？她是否使一个无辜的人失去了生命？**

我想，即使这个幸运（不幸）的精子和卵细胞已被选定，未来孩子的身份也因此确定，如果母亲一方在此时退出试管授精，大多数生命权的支持者也不会将其视作谋杀犯。这就意味着，未来孩子的基因并非关键因素。事实上，精子和卵细胞真正结合的那一刻，在道德上具有重大意义，“生命”从受孕开始。如果是这样，那么最重要的问题就是：“受孕一刻究竟发生了什么？”

嗯……那一刻发生了很多奇妙的事情。我不会在这里讲一堂生物课，不管怎样，我也不够格。对我们来说最重要的是，我们完全了解受精过程前后的细节，能够从分子层面上对其做出解释。我们知道精子的运动原理：精子内部中段的线粒体产生aTP（腺嘌呤核苷三磷酸），为精子尾部（鞭

毛）的运动提供能量。鞭毛中的动力蛋白将aTP中所含的化学能转化为动能，为精子尾部的运动供能，推动精子向前移动。我们也知道精子如何与卵细胞相遇：精子对卵细胞发出的一系列化学信号与热信号十分敏感。我们还知道精子中的幸运儿接触到卵细胞表面后会发生什么：卵细胞周围有一层糖蛋白膜，叫作透明带，其中含有化学感受器，能与精子头部的化学感受器进行匹配。在化学物质的相互作用下，精子分泌出一种消化酶，使自己能够穿过透明带，向卵细胞膜移动。随后，精子与卵细胞的细胞膜发生融合，触发一系列化学反应，阻止其他精子进入卵细胞。已经进入卵细胞的精子将所携带的基因物质释放出来，一层新膜会将该基因物质包裹起来，形成雄原核。同时，精子与卵细胞的融合也使得卵细胞的基因物质停止分裂，形成雌原核。两个原核在微管这层薄薄的聚合结构的牵引下相互靠近，相互融合。两组基因物质被拉到了同一个细胞核内，即受精卵的细胞核。至此，受精完成。之后，受精卵会分裂为两个细胞、4个细胞、8个细胞，直到形成一个空心的球体，叫作囊胚。囊胚进一步发展为原肠胚，形成外胚层、中胚层和内胚层三层独立的细胞。三个胚层将会分化形成不同的身体组织。例如，外胚层会发育为神经系统（脑和脊椎）、牙釉质以及皮肤的最外层（表皮）。

我将这些知识写出并非为了炫耀自己在发育生物学方面的知识，我也是查阅资料才弄懂的。我想要让你明白，我们对于生命发育的最初阶段的过程细节已经有了惊人的了解。事实上，与生物学家对这一过程一步一步、一个分子一个分子的了解相比，上述的概括根本只是沧海一粟。这一过程中当然也有未知的环节，但这些未知并非巨大的谜题，只是一些有待填补的漏洞。也许下一篇论文中，便会将一长串化学反应的下一个蛋白质描述出来。

对于人类发育过程如此精细的了解为生命权的支持者提出了一个严肃的问题。他们的观点是，受精的那一刻创造出了一个拥有生命权的人。受精在人类的发育过程中是一个奇妙的、决定性的瞬间。但据我们所知，这一过程并不神奇。事实上，人类卵细胞的受精过程与老鼠或青蛙的受精卵相比，没有丝毫更加神奇的地方。不论是受精过程，还是在发育的后期阶段，完全没有证据能够证明灵魂的产生。据我们所知，一切不过是有机分子在物理规律的作用下进行活动。

生命权的支持者该怎么办呢？他们可以坚持认为，受精的那一刻必然发生了某些神奇的事情，如果科学家们还没有发现这件事，那只能说明他们的无知，或者说明他们不够虔诚，存有唯物主义的偏见，但这种观点不过是一厢情愿，是没有任何证据支撑的空中楼阁。生命权的支持者会说，提出这种观点无异于选择权的支持者在毫无证据的情况下宣称，道德魔法发生于怀孕6～9个月期间，而不是发生在受孕的那一刻。

更加温和的生命权支持者也许会承认，我们确实不知道灵魂是在哪个阶段产生的，但鉴于我们对此一无所知，我们应当采取稳妥的方式。既然我们不知道灵魂何时产生，那么就应当禁止任何形式的堕胎。如果是这样，为什么要在受精这一步停止呢？为什么不能假设上帝为每个未受精的卵细胞都赐予了灵魂，精子只是提供了一些必要的分子呢？或者为什么不能假设上帝为精子赋予了灵魂呢？（给巨蟒剧团发个信号。）我们如何确定避孕措施不会杀死灵魂？为了稳妥起见，我们是不是应当禁止避孕呢？我们又如何确定禁欲行为不会杀死灵魂？为了更加稳妥，我们是否应当要求女性在子宫能够容许的情况下尽可能多地接受精子（也许承载着灵魂）？

当禁止堕胎的法令面临可能的例外情况时，生命权支持者的麻烦就更多了。2012年，共和党参议员候选人理查德·莫达克（richard Mourdock）声称，即使是强奸案例，他依然反对堕胎。他的解释在美国国内掀起了轩然大波。他说：

我认为，即使生命始于强奸这种可怕的场景，这也是上帝的旨意。

这句话传开之后，他的竞选过程便滑出了正轨。莫达克说上帝让女人被强奸！导致莫达克失手的这个问题比想象中更加严重，已经超出了堕胎的范畴。其实质是由来已久的“罪恶问题”，神学家们已经被这个问题困扰了几个世纪：如果上帝是全知全能的，他为什么允许像强奸（虐待儿童、校园枪击事件、大型地震等）这样的事情发生？面对这个问题的不只是莫达克一人，对所有信仰全知全能的善意之神的人来说，这都是一个问题。不论怎样，莫达克的言论在选民当中没有获得赞同，妇女的反对尤为强烈，他本人也未能在竞选中获胜。但我并不认为莫达克憎恨妇女，或是憎恨强奸案中的受害者。我认为他不过是想要做一名能够自圆其说的生命权支持者。他曾说过：

我相信生命始于受孕之时。我认为，只有在母亲的生命受到威胁的情况下，才能够实施堕胎。我自己也为此斗争了很久，但我逐渐意识到，生命是上帝的恩赐……

如果你真的相信“生命”始于受孕之时，并且相信上帝在那个时刻将灵魂附着在了生物物质上，那么说真的，我们何必要扰乱上帝这种抽象的赐予？对于生命权的支持者来说，莫达克的观点唯一值得质疑的地方就是他愿意为了挽救母亲的生命而实施堕胎。如果一位母亲的获救必须以牺牲她3岁的孩子为代价，那么我们可以这样做吗？

总之，生命权的支持者也许是正确的，也许我们真的拥有灵魂，也许上帝真的是在受精的那一刻将人类灵魂附着到了生物物质之上。但完全没有证据表明事实真的如此。与关于灵魂产生的其他理论一样，不论我们认为灵

魂产生于怀孕晚期，还是认为灵魂产生于受精之前，我们并没有更多的证据。当生命权的支持者自信地宣称胎儿拥有“生命权”时，他们与选择权的支持者一样，也在虚张声势，假装他们能够清晰地证明自己的观点。但事实上，他们拥有的仅仅是强烈的情感和毫无根据的假设。

关于堕胎的观点，有些能够引起人们的深切共鸣，有些则没有。有些观点甚至得到了正反两方的认同。如果人们对堕胎的态度众说纷纭，那么这些态度是从何而来的呢？和以往一样，从心理学角度稍作分析是大有裨益的。

也许你还记得第2章中的试验：婴儿们喜欢与帮助圆眼睛的圆形上山的圆眼睛小三角玩耍。也许你也记得，当圆形上不再有圆眼睛，孩子们看不到圆形自己努力滚动时，婴儿的偏好便消失了。没有了眼睛（所谓的“心灵之窗”）和自发的主动移动，圆形就只是一个形状而已。同样的，也许你还记得，一双眼睛便足以让大脑中的杏仁核警觉起来（图5.2），也足以让我们更加慷慨（图2.3）。眼睛能够触发人类的社会思维。

尽管眼睛的作用不可小觑，但试验发现，单纯的移动便能将没有特征的实体变为拥有心灵和思想的生物。20世纪40年代，社会心理学方面的先驱弗里茨·海德（Fritz Heider）与玛丽安·西梅尔（Marianne simmel）设计了一段著名的影片，确切地说，是由三种形状表演的默片。片中一个大的坏三角折磨另两个略小的形状，当两个形状逃跑的时候在后面追赶。影片中只有移动的形状，但人们会自动地为形状赋予目的（“那个大三角想要抓住他们。”“两个小形状想要逃跑。”）、情感、（“他们逃跑了，大三角很生气。”“成功地逃跑了，小形状很高兴。”）甚至角色特点（“大三角恃强凌弱”）。这个赋予的过程完全自发，人们甚至无法阻止自己进行这样的思考。人类能自动地看到社会活动，就像看到颜色与形状一样自然。

胎儿会动，胎儿也有眼睛。医学影像学产生之前，对堕胎的道德意义进行思考的人们大多将“胎动”作为道德上的转折点，因为胎儿从此开始产生人类可以觉察的运动。医学影像让我们在胎动之前就可以观察到胎儿的活动，甚至能看到胎儿的五官，比如眼睛。对很多人来说，这项技术将胎儿发育的神奇时刻大大提前了。

运动和眼睛会对我们产生巨大的影响，但这两个因素无法解释所有问题。我们多数人每天毫不犹豫吃下的动物也会运动，也有眼睛。*但与这些动物不同的是，胎儿已经多少有了人类的形状。它们拥有人类的手、脚，以及面部，他们的运动方式与人类也十分相似。毫无疑问，生命权的支持者总愿意展示胎儿的照片，特别是对手、脚以及面部的特写照片，便是出于这些考虑。因此，1984年上映的支持生命权的电影《无声的呐喊》（*The Silent Scream*）在人群中引起了轰动。这部电影与海德和西梅尔所制作的

影片有种怪异的相像之处。电影记录了超声波观测下的堕胎过程。我们可以看到胎儿向子宫深处移动。旁白解释说，胎儿“有意识地”远离堕胎设备，它在“焦躁地”、“剧烈地”移动。在影片中最关键的一刻，面对伸过来的流产吸引器，胎儿张开了嘴。下一刻，胎儿的头部便被压碎，以便从子宫颈口排出。不论你支持生命权还是选择权，观看这部影片都让人感到难以接受，这也正是影片希望达到的目的。《无声的呐喊》唤起了大脑的自动模式，提出了反对堕胎的有力“论据”，任何实在的（手动模式）论据都无法与之相比。

《无声的呐喊》之所以会获得成功，是因为胎儿看上去与人类十分接近。如果影片拍摄的是怀孕的前三月内进行的流产，那时候，发育中的胎儿不过是一团细胞，对人的冲击也不会如此之大，因为破坏一团细胞似乎并不是多么可怕的事情。这里便产生了道德家们直觉上的两难困境。在胎儿显出人形之前，堕胎并不会让人感觉很糟糕；而在胎儿完全发育成婴儿的形状之前，堕胎也并不会让人产生糟糕透顶的感觉。但胎儿究竟何时完全发育成婴儿的形状，又在何时显出人形，这之间并没有明确的分界线。粗略来说，人类胚胎发育初期与老鼠或青蛙的胚胎并没有明显的分别。这个阶段与胚胎开始显出人形的阶段之间也同样没有明显的分界线。整个过程中，唯一具有明显界限的事件就是受精过程。但在受精卵阶段，发育中的胚胎完全无法唤起自动模式的情感。很显然，那不过是一团有机分子而已。如果我们不对堕胎加以限制，看上去与婴儿一样的生命（某人！）就会被扼杀。但如果我们禁止堕胎，那么就是强迫自由的人们为了一团分子而严重扰乱自己的生活。在两个极端之间，我们也无法找到能让心灵感到安宁的中间点。

那么我们该怎么办呢？很大程度上来讲，我们只是沿用部落内其他成员的做法。多数部落都相信灵魂的存在，由于种种原因，这是一种非常自然的信仰。如果你坚信人类拥有灵魂，你就必须相信，灵魂是在某个时刻产生的，最可能的时刻便是受孕之时。的确，一个单独的细胞看上去不像是拥有灵魂的一个生物，但你还有别的选择吗？受孕之前，两个细胞彼此独立；受孕之后，又找不到一个转折性的事件。至今为止，受精是最有可能产生灵魂的时刻。因此，如果部落的长者告诉你这便是“生命”开始的一刻，而你又无法提出更加合理的理论，你便应当接受这种说法。更重要的是，提出反驳意见会让你像个“外人”。

相反，如果你的部落不相信灵魂的存在，或者允许人们自行思考灵魂问题，你该怎么办呢？根据某些人对灵魂产生时刻的推测做出自己的决定显然不是个好办法。如果你的部落看重个人选择，这一点便显得尤为重要，不论是针对堕胎还是更广泛的问题。但并非所有事都是无所谓的。也许你的部落对灵魂的产生没有明确的态度，但至少人们肯定：杀死婴儿是绝对

不允许的。因此，为了稳妥起见，我们也不能允许人们杀死形似婴儿或可能成为婴儿的生命，即能够在子宫之外生存的一切生命。糟糕的是，形似婴儿的界限十分模糊。早在发育初期，胚胎便能够自主地活动，拥有与人相似的手、脚与脸庞。我们该怎么办呢？

支持选择权的人们内心必须有一种奇怪的、平衡的情感。即使是选择权的支持者，也几乎没有人能够坦然面对将人形的生命杀死，很多人甚至看到杀死形似动物的生命时也会感到不适。但选择权的支持者也同样不愿对别人发出指令，特别是对女性的命令。因此，选择权的支持者不得不在“不能命令别人做事”与“不能杀死拥有这样外表的生命”之间挣扎，找到令人不适但却无法避免的平衡。

这一切对于堕胎的争论有什么意义呢？这意味着，我们所有人几乎都在虚张声势。我们自信满满地维护“生命权”和“选择权”，这都不过是手动模式下的虚构症。我们拥有的不过是不成熟的直觉理论，受我们自己也不甚了解的认知机制驱动。我们试图为直觉理论蒙上一层理性的面纱。一旦除去这层高尚的权利主张，便所剩无几了。一个诚实的生命权支持者应当是这样的：

我相信人的躯体中驻有灵魂。尽管我没有真凭实据，但我相信这一点，我信任的所有人也都相信这一点。我不知道灵魂是如何进入身体的，但我所信任的人说，当精子与卵细胞相遇时，新的灵魂便产生了。我不知道这一过程具体是怎样的，但我也提不出更好的观点了。我猜，从受孕一刻开始，就有一颗人类的灵魂开始了。扼杀一颗无辜的人类灵魂是不对的。我知道这种观点很大程度上取决于信仰。我也理解我们应当尊重彼此的信仰。但我真的无法眼看着人们扼杀可能拥有灵魂的生命，不论这条生命多么渺小。对于不愿怀孕的人来说，我知道这是很困难的。但这些人选择进行了性行为（强奸案例是特殊情况，不在此列）。而扼杀可能拥有人类灵魂的生命并不是对自己的选择反悔的合法方式。这就是我的想法。

一个诚实的选择权支持者则应这样想：

我相信人们应当自由地独立思考，做出自己的选择。同样的态度也适用于堕胎，至少在怀孕早期应是如此。怀孕的前3个月，胎儿看上去有些像人，但其实它看上去与青蛙也很像。尽管我不愿杀死青蛙那么大的小人，但我认为，强迫女性完成她们所不愿意的妊娠过程是更加糟糕的事。我知道有很多人愿意收养婴儿，但将自己生的婴儿送给别人是极其痛苦的事情。强迫女性经历这一过程对我来说，比杀死一个青蛙那么大的小人更难以接受。然而，怀孕6～9个月时，胎儿不再像青蛙一样大，而是发育出了婴儿的形状。杀死一个婴儿显然是不对的。所以如果你腹中的胚胎看上去还像青蛙那么大，我认为你有权选择放弃它的生命。但如果你腹中的胎儿

已经发育出了婴儿的形状，而不再是青蛙大小的生命，那么我认为你别无选择，必须让它生存下来。这就是我的想法。

关于堕胎的争辩归根结底便是这样：支持双方的是强烈且复杂的情感，我们既无法为其找到依据，也无法直接忽略。那么，作为现代道德牧民的我们，应当何去何从呢？

## 堕胎：实用主义的解决方案

既然我们认为“权利”的概念只是双方的虚张声势，那么我们就可以开始以深度实用主义思想分析问题了。我们不再试图探究“生命”从何时开始，而是从另外一系列问题开始思考：如果将合法堕胎的途径封死，会发生怎样的情况？反过来，情况会怎样？这些政策会对我们的生活产生怎样的影响？这些都是复杂而实际的问题，想要回答并非易事，但我们可以根据已有信息进行合理的推测。

如果堕胎被法律所禁止，人们行为的改变大致可以分为三类。第一，有些人会改变性行为方式。有些人可能会完全放弃性行为，至少在一段时间内不会恢复。有些人可能会降低性行为的频率，还有人可能会采取更多的方式降低怀孕的可能性。第二，有些人会通过违法途径或者出国等其他渠道进行堕胎。第三，有些人会不得已将婴儿生下。他们可能会将婴儿送给别人收养，也可能会亲自抚养。

这些情况加在一起会怎么样呢？我们先来讨论第一种情况。对多数成年人来说，不以繁衍后代为目的的性行为是生活中令人极其愉快、极其享受的一部分。除去年轻人与躁动不安的人们，在一夫一妻制度下，对于婚姻生活稳定的夫妻来说，不以繁衍后代为目的的性行为是幸福感的一个重要来源。有生育能力的夫妻可以通过避孕措施进行安全的性行为。但我们都知道，即便采取了正确的避孕措施，也仍有可能发生意外。因此，对上百万对经常进行性生活的夫妻来说，堕胎的选择是对抗意外怀孕的一层重要屏障。

在实用主义账单的另一侧，有些性行为确实是有害的。如果堕胎被法律禁止，有害的性行为可能也会减少。比如双方自愿的会对情感造成损害的性行为、感情上尚未做好准备的青少年之间的性行为、乱伦、强奸等都属于有害的性行为。如果人们由于恐惧怀孕而减少性行为，也许通过性行为传播的疾病也会因此得到控制，这也是一个好的连带作用。但我们所不能确定的是，禁止堕胎是否会大幅度减少有害性行为。强奸犯似乎不会因为知道受害者无法堕胎就终止自己的行为。当然，禁止堕胎能够减少青少年性行为。但我们依然无法确定这样的结果最终是好还是坏。也许青少年只有充分了解到自己行为的后果，才算真正为正常的性生活做好了准备。

总之，考虑到禁止堕胎对人们性行为方式的影响，数百万对性生活正常的夫妻将会蒙受损失，生活的幸福感会大幅降低，并且无法得到明确的补偿。

接下来要考虑的是堕胎的其他途径。对于充满智慧的人类来说，在法律上禁止堕胎只会使堕胎费用变得更加昂贵，堕胎途径更加不方便。为了迎合人们非法堕胎的强烈需求，国内必然会产生相应的市场。不幸的女性只能到这种市场上寻求帮助。我在这里并不想描述非法堕胎的恐怖画面。从功利主义的角度来看，迫使人们寻找其他出路实施堕胎必然会导致糟糕甚至悲惨的后果。

最终，我们来考虑出生人口增加所造成的影响。强迫妇女经历怀孕过程是可怕的行为。即使在最好的条件下，怀孕也是对女性情感的巨大挑战。迫不得已怀孕的女性也许会对胎儿疏于照顾，也许是有意为之，也许是无心之举。怀胎十月直到生产不仅是对女性情感的巨大挑战，还有可能严重扰乱一个人的生活。总之，强迫女性违背自己的意愿，将孩子生出来是非常糟糕的事情。

不管怎样，有人也许会说，强迫女性完成妊娠过程所带来的积极影响更大。通过生产，女性使一个新的人得以开始生活。如果这位女性不愿亲自抚养婴儿，她可以放弃婴儿，送给别人收养。如果运气够好，这名婴儿将会进入一个充满爱的家庭，拥有充足的资源。我们很难描述孩子生母所承受的苦难，不论有多么煎熬，都比她亲生孩子的生存更加重要吗？当然，不幸的是，并非所有等待被收养的儿童都能够找到收养家庭。一旦堕胎成为违法行为，那么可供选择的好的领养家庭便会越来越少。即使孩子被收养之后的状况远远不尽如人意，我们也不能宣称母亲的伤痛和苦难高于孩子的感受。只要孩子被收养之后过上了有意义的生活，我们就不能说生身母亲所遭受的痛苦比其亲生子女的生活更加重要。

在有些情况下，母亲，也许和（或）父亲会决定亲自抚养孩子。很多（多数）情况下，事情的进展会比较顺利。很多幸福家庭中的孩子起初都是意外怀孕的结果，父母最初也曾有过放弃的念头。有时候，这个不受欢迎的孩子的生活也许并不十分顺利。但除非这个孩子的生活完全没有意义，或者这个孩子的存在会使世界整体变得更糟，或者从更加现实的角度来说，如果这个孩子的存在会妨碍另外一个孩子的幸福生活，或是会妨碍另外一个孩子将整个世界变得更加幸福，否则堕胎便不是更好的选择。

只有采用这种奇怪的功利主义计算方式，生命权的支持者才能够有力地证明自己的观点。如果堕胎成为非法行为，世界上会多出很多人。有时候，这些增加的人口会使世界的幸福感总值下降。但总体来说，我们无法宣称禁止堕胎后多出生的人口会变得不幸福，或者会让整个世界更加不幸福，

对此我们毫无信心。当然，这是一个复杂的实际问题，很大程度上取决于是否能找到足够多的优秀收养家庭。如果优秀的收养家庭足够多，我们便很难宣称流产是对胎儿或婴儿最好的选择，或者说整个世界会因为它们的消失而变得更加幸福。

我们能够得出怎样的结论呢？我随手记下的要点是这样的：一方面，禁止堕胎将使上百万人失去一个重要的安全保障；会促使一些富人花高价进行堕胎；还会迫使一些绝望的妇女和女孩儿铤而走险，进行非法堕胎。禁止堕胎还会扰乱很多人的生活计划，让他们在尚未做好准备，或完全不想要孩子时迎来孩子。这些代价都十分沉重。另一方面，禁止堕胎会使很多本来没有机会生存的人拥有生命。基于优秀收养家庭的数量多少以及其他一些因素，这些人的生活有可能会很好。那么这些分析能够得出怎样的结论呢？我们是不是又一次陷入了僵局？

我不这样认为。这个从功利主义角度出发的论据很好，对生命权的支持者来说可谓雪中送炭。但问题在于，这个论据太好了。也许你还记得之前的讨论，我们提到，避孕措施、禁欲思想与堕胎一样，都会造成生命的消失。如果我们以堕胎剥夺了人的生命权为理由而反对堕胎，那么我们也应当反对避孕、反对禁欲。因为这两种行为产生的效果与堕胎完全相同。然而，几乎不会有生命权的支持者愿意采取这样的立场。

事实上，生命权支持者这种深层次的论证与功利主义对极端利他主义的论证，以及将自己变成幸福之源的论证十分相似。将自己变为幸福之源的一种方式就是更加有效地分配资源，牺牲富人，帮助穷人。但我们还有另外一种选择，那就是感染带动更多的幸福之人。还有更好的选择，那就是培养更多幸福的小功利主义者，他们会愿意为了他人的幸福而努力奋斗。当然，这个想法并非毫无道理。但对于没有英雄情结的人来说，这个要求实在太高了。如果我是拥有造人能力的神明——创造的物种可以让尽可能多的同伴感到幸福，或者会对这种做法感到迟疑——如果其他一切条件完全相同，我会选择让人幸福的物种。我们不愿创造出更多幸福的人类，我想这一点与道德无关。所有活着的人类达成了一种密约，用以对抗那些无法表达自己意见的人们，事实上，他们其实是没有代表的多数派。由于我们自私的选择，这些假设中的无助的多数派根本没有机会对自己的不存在进行抗议。

哦，真是一件憾事。不论怎样，我们不能过分认真地看待生命权支持者的功利主义论据。但选择权支持者的功利主义论据并不是太好，只是一般好而已。扰乱人们的性生活，扰乱人们的生活计划，迫使人们去国外或寻求非法途径堕胎都是非常糟糕的事情，会使很多人的生活更加不幸，甚至会缩短很多人的寿命。最终，站在深度实用主义的角度，这就是我支持选择权的原因。我并没有使用“权利”为自己辩护，只是现实地考虑了各种后

果。

如果你是一位诚实的生命权支持者，不愿用“权利”来虚张声势，那么你面临着两种选择。第一，你可以直率地提出部落中深奥的信仰，板着脸坚持宣称这种信仰是整个世界的运转法则。但如果你这样做，选择权的支持者可能会提出这样的问题：“上帝是在精子的头部接触到透明带时将灵魂创造出来的吗？或者说灵魂产生于精子接触到细胞膜的瞬间？精子的全部基因物质进入到卵细胞就够了吗？或者上帝会等到雄原核与雌原核融合之后再行动？灵魂的产生过程与融合过程重合吗？还是融合过程进行到一半时灵魂才产生？”这个时候，生命权的支持者便不得不承认，他们无法拿出证据回答这些问题。但即使这样，他们依然坚持，这种基于信仰的答案揭示了世界的规则。

与此相反，对选择权的支持不必建立在没有事实依据的理论主张上，也不必依赖经不起认真推敲的证据。选择权的支持者们确实尚未找到一条原则，划清界限。但这种情况也许是不可避免的。不论引发人们道德思考的是什么因素，这些因素并非一下子出现在某个神奇的时刻。既然我们无法相信神奇的时刻，选择权的支持者就必须找到一条界线，同时也不得不承认，他们所找到的这条界线具有一定的随意性。到目前为止，也许我们找不到更好的地方划定这条界线。但基于通用货币，如果有人能提出更好的观点，将界线划在别的地方，深度实用主义者一定会虚心接受的。

## 等待戈多

在堕胎这个问题上，也许你认为这个实用主义、功利主义的“出路”并不令人满意。事实上，我们确实没有那种找到正确答案的感觉。这种暂定的支持选择权的结论看上去并不像是一场胜利，倒像是一份无限期的停火协定，尽管协定内容明显倾向于其中一方。如果你对这种感觉不甚满意，也许会试着寻找一场真正的道德胜利。也许你会坚持寻找一种支持堕胎的理论，既能够自圆其说，又让人感觉安心。事实上，每当我们进行道德探究时，想要的都是这样的感觉。我们放弃得是不是太早了？

很多道德思想家都会做出肯定的回答。很多生物伦理学家都曾经尝试解决这个我没能解决的问题，试着在面临生死抉择时，找到直觉上说得通，不依靠功利主义思维的对错判断标准。在堕胎和生物伦理学之外，很多道德哲学家致力于设计一种复杂的道德理论，从而代替19世纪的功利主义优秀传统。难道这些人都是在缘木求鱼吗？我想是的。我无法证明这一点，也不想试着证明。事实上，这一部分中，我想要解释自己为何对复杂道德理论的发展前景抱持消极态度。*（如果你对复杂道德理论为何不可能成功不感兴趣，就可以跳过这个部分了。）

一切都要从道德思维的双加工理论说起。我们想要的是手动模式的道德理论，要能够用文字清晰地表达出来，并且总是能够与大脑的自动模式达成一致意见。如果大脑的自动模式认为杀死6~9个月大的胎儿是错误的，而杀死三个月内的胚胎则是可以接受的，那么我们需要的道德理论就是赞同这种直觉，并能够做出解释的理论，依此类推。简而言之，我们需要的道德理论要能够将直觉梳理清晰，并为之找到理由。用罗尔斯的话说就是，我们想要找的是一种“反思平衡”，使道德理论能够与大脑中“经过深思的判断”相契合。

但我们的直觉本身便没有条理性可言，也并非为了达到真正的道德目的而存在。自动模式采用的是启发式思维，这种有效的算法在多数情况下都能够得出“正确”的答案，但偶然情况也会出现意外。我之所以将“正确”一词放在括号里，是因为即使大脑的自动设置按照预设的情况正常工作，也并非必须表现出真正“正确”的道德观。有些直觉只不过反映了传播基因这项生物学责任，可能会使我们优先考虑自身以及本部落的利益，而非外人的利益。了解到这一点，我们便可以开始试着理清思路了。梳理道德直觉之前，我们可能需要首先抛弃所有的偏见。如果我们使用科学的自我认知揭露了所有的偏见，最后结果会是怎样的呢？

我相信我们会得到类似功利主义的结论。为什么呢？首先，我在第8章中提到过，功利主义是合乎逻辑的思维，不仅你我可以理解，只要不是精神病患者，大脑的手动模式能够正常工作，所有人都能够理解。反对功利主义的观点中，唯一有威胁的观点是，某些情况下，根据功利主义思维得出的结论在直觉上是错误的，在假设出来的场景中尤为明显。第9章与第10章中，我们审视了很多例子，逐渐对反对功利主义的道德直觉生出了怀疑。这些反对功利主义的直觉似乎总会对一些与道德无关的因素非常敏感，比如，动作的执行是由手推的动作完成还是通过扳动开关完成。我希望以后能够找到更多类似的例子。

其次，我在考虑：如果反对功利主义的道德直觉转而对关乎道德的因素十分敏感，这意味着什么呢？一种可能性是，直觉判断会促使事件达成好的结果：对暴力等造成糟糕结果的事件，直觉会产生负面反应；而对于帮助他人等造成好结果的事件，直觉会产生正面反应。（也就是说，说一不二的道德直觉成了“规则功利主义”。）如果这就是我们得出的结论，说明自动模式是不完美的功利主义装置，那么功利主义的立场便得以进一步巩固。如果触发自动模式反应的关乎道德的因素与好的结果无关，又会是怎样呢？很自然，我们会认为直觉会受到权利概念的支配。比如，将人推下天桥的行为在直觉看来是错误的，这种感觉反映出的也许是这个人的权利遭到侵犯这个事实。但倘若没有独立的、非功利主义的权利理论（根据不证自明的道德公理形成的理论），我们如何能证明这种感觉是正确的呢？

我们怎么知道究竟是直觉受到权利概念的支配，还是直觉编造出了“权利”这个幻象呢？对于权利的完整描述（非功利主义的）究竟是什么样的？

你也许会在某个时刻突然醒悟：一代又一代的哲学家与神学家心中的道德也许并非真正的道德。道德并非零星分散的抽象真理，有限的人类思维也许根本无法望其项背。道德心理学闯入道德哲学的抽象领域也绝非偶然。事实上，道德哲学是道德心理学的表现形式。与更加宏大深邃的心理学与生物学相比，道德哲学不过是露出水面的冰山一角。意识到这一点后，你的整个道德观便会发生变化。数据和事实都表现出了不同的含义，相互矛盾的道德哲学不再是抽象哲学世界中的散点，而是人类大脑双加工机制的必然后果。

西方道德哲学主要包括三个思想学派：功利主义/结果主义（边沁与密尔）、义务论（康德），以及德行伦理学（亚里士多德）。从本质上讲，这三种思想学派是手动模式为自动模式的决定解释原因的三种不同方式。我们可以使用手动模式思维清晰地描述自动模式思维（亚里士多德），可以用手动模式思维为自动模式思维寻找合理解释（康德），也可以用手动模式思维打破自动模式思维的局限（边沁与密尔）。有了这样的印象，我们可以从心理学角度对西方道德哲学进行简要审视。

假如你是一个部落的首席哲学家，事情会是什么样呢？在部落内部可能存在很多道德争端，但这些争议根本上都是“我”与“我们”之间的矛盾，或是“我”与“你”之间的矛盾，也就是公地悲剧中反映的问题。在部落内部，只有“我们”的存在，因此并不会出现真正的道德争议，也不存在“我们”与“他们”的价值观冲突。因此，作为部落的首席哲学家，你的任务并不是调解相互矛盾的道德世界观，也不是对部落内部的常识提出质疑。相反，你的任务是将这种常识提炼为规则，编纂成典，为部落传统智慧的积累提供资源。你的任务是不断重复人们已经了解，但有时会忘却的常识，并给予提醒。

在众多西方哲学家中，亚里士多德在常识方面可谓无人能及。与他的导师柏拉图不同，亚里士多德没有提出任何激进的道德观点，也没有提出任何定式。对亚里士多德来说，道德上乃至其他层面上的善行都是一种复杂却平衡的行为，其最好的表达方式就是美德，还有促人进步的持久的习惯和技能。例如，亚里士多德认为，面对危险，我们既不能鲁莽也不能退缩，我们需要的是勇敢，这种美德便是两种有害的极端中间的平衡位置。与爱、友谊、工作、玩耍、冲突、领导等词语相关的美德都需要人们自己做出平衡的行为。对亚里士多德来说，世间并没有清晰的一套规则，能够告

诉人们如何达到善的平衡。一切都要在实践中证明。

作为一名伦理学家，亚里士多德从根本上讲是一位部落哲学家。阅读亚里士多德的著作，你便能感受到一位睿智温和的贵族，从古老的马其顿和雅典走来。古老的马其顿和雅典贵族所接受的一些教育至今依然适用，你也可以从中学到一些做人的道理。但亚里士多德不会帮助你思考堕胎是否错误、是否应当捐出更多的钱来帮助远方的陌生人、发达国家是否应当采取统一支付的医疗体系等。亚里士多德的哲学体系以美德为基础，像一位慈祥的祖父那样提出建议，这样的体系本身便不是为了回答这些问题而存在的。我们无法借助美德解决部落争端，因为一个部落视为美德的行为在另一个部落也许是一种罪恶。即使事实并非完全如此，在两个部落发生争端时这种情况也会发生。

在现代道德哲学家中，亚里士多德的美德理论重新得到推崇。*为什么会这样呢？启蒙运动的最大愿景是希望哲学家们能够构建出一种系统性、普适性的道德理论，即元道德。但结果如我们所知，哲学家们没能找到让人感觉正确的元道德。（大脑的双加工机制决定了这个任务是不可能完成的。）面对失败，我们可以继续尝试，也可以就此放弃。并非放弃寻找元道德，而是不再执着于感觉正确的元道德。（这也是我的建议。）当然，我们也可以将启蒙运动提出的任务完全放弃。我们可以宣称，道德太过复杂，无法用清晰的一套规则进行界定，我们只好通过实践来磨炼自己的道德情感，追随道德模范的脚步。面对人类价值观这团乱麻，支持亚里士多德观点的现代哲学家只能不断祈祷，毫无益处。

总之，亚里士多德与同侪十分擅长描述在某个部落当中如何成为优秀的一员，各种部落的成员都能够从中受益。但面对部落之间发生的争执，面对现代道德问题，亚里士多德几乎无能为力。因为某个部落眼中的美德可能恰是其他部落眼中的罪恶。（若想放弃启蒙运动提出的任务，我们还可以把自己变成“相对主义者”或“虚无主义者”，不再试图承认某一个特定部落的常识，而是宣称所有部落的认知都不正确，引入一种敷衍了事的态度。）

如果你不愿满足于亚里士多德（或现代“相对主义者”）的理论，你可以试着用自己的手动模式思维描述本部落的道德观，并且为其找到合理的解释。你可以试着证明本部落的道德原则与数学定理一样，是放之四海而皆准的。这便是伊曼努尔·康德的思想。

数学家们努力证明各种定理，但他们从未想过要证明所有定理。他们选择看似正确或是可能正确的数学表述，然后根据基本原理和公理试着推导。数学家们在这方面取得了巨大的成功，从毕达哥拉斯定理到安德鲁·怀尔斯（andrew Wiles）对费马大定理（Fermat’s last Theorem）的证明等。既

然如此，为何不能以同样的方式解决道德问题呢？哲学家们为何不能从基本原理中推导出有趣的道德真理呢？

这就是康德的愿望，也是此后很多哲学家们的雄心壮志。他们希望能够论证本部落道德的正确性，尼采将其称为康德心中的“秘密玩笑”。但这对康德来说是一件尴尬的事吗？数学家们努力证明看似正确的数学论述，这丝毫不是丢人的事情。那么为何康德的愿望就成了秘密的玩笑，而不是高尚的事业呢？

康德的雄心壮志本身并没有问题，问题在于，康德不愿面对失败。数学家们已经成功地证明了无数个数学争议，但通过对基本原理的推导证明，从未有任何道德争议得以解决。康德想要证明自己道德观点的正确性，这种心理过于迫切，导致他忽视了论据中存在的缺陷。也就是说，康德越过了界限，从逻辑推理跳到了合理化思维，难怪尼采会在背后偷笑了。

如果我们不用同情的眼光看待康德，便不难发现，他的论据是站不住脚的。例如，他认为手淫之所以错误，是因为在这个过程中人将自己作为了工具。（果真如此吗？那么如果某个人为了舒服而按摩自己的胳膊，也是错误的行为吗？）康德关于手淫的观点并不是他的得意之作，但据我所知，他最著名的观点也不过如此。例如，康德有一个著名的观点，认为说谎、违背承诺、偷盗以及杀戮之所以错误，是因为这些行为的“标准”无法得到“统一”的界定。如果每个人都去说谎（或违背承诺），那么讲真话（或遵守承诺）的习俗便会遭到破坏，那时便不再有所谓的说谎和违背承诺了。同样的，如果每个人都去偷盗，那么个人财产这个概念便会淡化，也就不再有偷盗的概念了。如果每个人都去杀戮，那么最终便会到达无人可杀的境地。

这些观点都十分巧妙，但它们都极度缺乏证据的支撑。一方面，仅仅因为一种行为不能按照康德的标准得到统一的界定就判定它是错误的，不论从逻辑上还是直觉上都无法服人。以时尚为例，如果所有人都变得时尚起来，那么也就无所谓时尚了，全民时尚也会削弱时尚的概念。但我们并不认为时尚是不道德的行为。同样的，一些恶劣的行为并不会削弱自身的概念。比如打人，我们甚至可以互相殴打，直到时间的尽头。这是非常糟糕的事，但并非完全不可能，而不可能恰恰是康德观点的关键。

康德的追随者们完全了解其观点中的瑕疵，并对此给出了各自不同的解释。但我们至少可以宣称：在将近两个半世纪的时间里，尽管不乏努力，但还没人能够成功地完善康德的观点，提供严密的道德证据。也没人能够成功地证明一条实质性的道德观点。我的意思是，没有任何道德争端曾经通过证据得到解决。当然，很多非常睿智的人都曾尝试过。其中最著名的当属约翰·罗尔斯（John rawls）。在其著作《正义论》（*A Theory of*

*Justice*）中，罗尔斯试图证明，他所推崇的平等主义式的政治自由主义可以由几个最基本的假设推导而来。我不知道是否有人相信罗尔斯的论据能够真实支撑其结论，但很多人相信，罗尔斯为非功利主义的道德和政治哲学开了一个好头。对此我仍持怀疑态度。事实上，我认为罗尔斯在《正义论》中提出的中心论点与康德一样，其本质都是合理化思维。***

我们可以对本部落的自动模式思维进行描述（亚里士多德），也可以试图证明其正确性（康德）。然而，两种哲学手段都无助于解决我们所面临的现代道德问题，因为一切问题的罪魁祸首就是我们的部族思维。因此，我们的唯一出路就是打破自动模式思维的局限，将问题（全部）转入手动模式思维。我们应当放弃部落内部的道德敏感性，也不再试图将其合理化，我们应当在通用货币的系统中，试着从共同的价值观中寻求共识。

亚里士多德以及拥有类似观点的学者也许是正确的。也许世上确实存在唯一的一套道德美德，应当被奉为所有人的追求。或者说，康德以及拥有类似观点的学者也许是正确的。也许世上确实存在真正的道德理论，能够通过基本原理推导而得。或者从更加谨慎的角度来说，我们也许可以将人类混乱的价值观进行整合，形成一个更加清晰、复杂的道德理论，对直觉的是非判断进行更加清晰的描述。也许吧。但在等待戈多的同时，我还想提出一种更加实用的方案：我们应当简单一些，让世界变得尽可能幸福。这种观点并不能解决我们的所有问题，但至少现在，这是现代牧民面前最好的选择。

## 我为何采取自由主义立场，我在何种情况下会改变自己的想法

我是一名大学教授，住在马萨诸塞州剑桥市。无须任何社会学调查便可以判定，我是一名自由主义者。（我所谓的“自由主义”是美国含义下的自由主义：偏左翼，与自由意志论者和某些“古典自由主义者”相比，对积极的政府行为持更加温和的态度。）我的自由主义观点完全在意料之中，但这种观点是合理的吗？我所在的自由主义部落是否与其他部落一样，拥有自己的直觉判断和成套的合理化解释呢？

某种程度上来讲，是这样的。上文中我们看到，在关于堕胎的讨论中，自由派也提出了毫无根据的主张，提供了自相矛盾的论据。但在我看来，自由主义部落与其他部落有着本质的区别。世界上存在着很多的传统部落，成员们拥有共同的历史，拥有一系列共同的“专有名词”（神灵、领袖、文本、圣地等）。然而今天，世界上存在着两个元部落，都属于后部落时期的部落。其中的成员并没有共同的历史，也没有共同的专有名词，而是因着共同的抽象理想而相聚在一起。一个元部落便是我所在的自由主义部落。我过去不总是支持自由主义的，因此未来某天，我也许还会再次对其

提出反对意见。现在我成了自由主义者，因为我相信，在现实世界中，自由主义部落所采取的政策会使世界的幸福感更高。但我在骨子里并非自由主义者。我首先是一名深度实用主义者，然后才是自由主义者。如果你有足够充分的证据，便可以说服我脱离自由主义。

为了理解自由主义部落的特殊性，我会把自己对道德与政治的理解与乔纳森·海特的观点进行比较。前文已经对他的观点进行了介绍，我本人的思维也深受其影响。第1、2、3章就道德的整体进化和心理学方面进行了分析，海特和我对此持赞同观点。前三章的主要内容可概括如下：

> 道德是经过生物学进化和文化发展而得来的一整套心理能力，为了促进合作而存在。（第1章）
>
> 从心理学层面来看，道德主要通过情感道德直觉实现。情感道德直觉是一种本能反应，使我们对某些人的利益加以重视，并鼓励他人抱有相同态度。（第2章）
>
> 不同的人类族群拥有不同的道德直觉，这便是重大分歧的来源。分歧的产生一方面是由于不同族群看重的价值观不同，另一方面则是由于自利性偏差，包括潜意识里的偏见。当人们的意见产生分歧时，往往会利用推理的能力为自己的直觉判断寻找合理化解释。（第3章）

除了这些科学的描述之外，海特和我至少在另外一条规范性描述上也达成了共识。这项描述是海特的著作《正义之心》（*The Righteous Mind*）的核心思想，我将其概况如下：

> 为了融洽相处，我们应当不再那么自以为是。我们应当认识到，几乎所有人都是好人。冲突之所以产生，是因为我们分属不同的文化群体，拥有不同的道德直觉。我们十分擅长揭穿对方为自己找出的合理化解释，但我们在揭发自己这方面做得还不够。具体来说，自由派与保守派应当试着互相理解，脱下虚伪的面具，乐于相互妥协。

这些都是非常重要的观点。但不幸的是，它们对我们的帮助也仅限于此。思想开放与放弃自以为是也许能够促进道德问题的解决，但其本身却并不能算作解决方案。

我与海特在观点上的第一个重大分歧在于推理或是手动模式在道德心理学中扮演的角色。我相信手动模式思维在人们的道德生活中起到的作用至关

重要，是我们的第二道德罗盘。但海特对此持不同意见，他认为道德推理在道德生活中占据的地位是次要的。他写有一篇著名的论文《情感是狗，推理是尾巴》（*The Emotional Dog and Its Rational Tail*），题目便已将他的观点清晰地表达了出来。（需要声明的是，海特并不赞同以这种比喻概括自己的观点。**）稍后我们会回到道德推理的问题上来。眼下我们需要考虑的是，放弃自以为是、摘下虚伪的面具为何不足以解决我们所面临的道德问题。（海特也同意这一看法。）

再次以堕胎的问题为例。有些自由主义者认为生命权的支持者厌恶女性，想要控制女性的身体。有些社会保守主义者相信，选择权的支持者是不负责任的道德虚无主义者，对人类生命缺乏基本的尊重，代表了“死亡文化”的一部分。像这样出言不逊的部落道德学家随处可见，而海特的观点提出得正是时候。可是那又能怎样呢？假设你是一位支持自由主义的成年人。你十分明白，生命权的支持者是站在道德的总体层面考虑问题的，他们并没有恶意，也不是疯子。即使是这样，你会出于相互妥协的考虑，同意为堕胎设置种种额外限制条件吗？同样的，成熟的自由主义者应当本着相互妥协的考虑，不再为同性恋伴侣争取完全的公民权利，而是致力于为他们争取更多的公民权利吗？对于思想开放的自由主义者来说，他们是否应当争取严格但又不太严格的环境法规来减缓全球变暖呢？当然，成熟的社会保守主义者也面临着同样的问题：他们是否应当“理性分析”，即使他们认为在怀孕早期堕胎等于谋杀，也要在这一问题上做出让步吗？

承认对方并非心怀恶意是一回事，但承认对方正确，或部分正确，或者承认对方的信仰和价值观与自己的同样有道理却是另外一回事。对自己的观点不再盲目自信是重要的第一步，但最重要的问题依然没有得到解答：我们应当相信什么？我们应当何去何从？

海特提出了一个更加具体的理论，解释了自由派和保守派产生分歧的原因，叫作道德基础理论。海特认为，自由主义者极度缺乏道德敏感性。海特定义了六类“道德基础”，每一类都包含了积极和消极的词汇：关心与伤害；公平与欺骗；忠诚与背叛；权威与颠覆；圣洁与堕落；以及新加入的自由与压迫。每一类基础都有其相应的一套道德情感。例如，关心与同情相关；圣洁与（对圣洁之物的）敬畏和（对不圣洁的污秽之物的）厌恶相关。海特将这些道德情感上的感受与舌头上的5种味觉感受器相比较。甜、咸、酸、苦、辣5种味道分别对应舌头上5种不同的味觉感受器，同样的，人类的情感思维中包含6种相互独立的道德感受器，能够对分属于6种道德基础的行为或事件做出相应的情感反应。例如，一位身处苦难之中的孩子会唤起道德思维中的关心与伤害感受器，让人产生同情之心。最重要的是，不同的文化群体（此处指代部落）拥有不同的道德味蕾，因此产生了不同的道德品位。海特认为，自由主义者的品位极差，他们能够感受关

心、公平与自由，但却无法“尝出”忠诚、权威和圣洁。

对于这个“六类理论”，我本身还是存有疑虑。但海特的理论中有一个重要方面看上去是十分正确的，并且拥有完善的证据作为支撑：**有些道德价值观是自由派和保守派同样看重的，但有些则并非如此。海特向其研究助手问道：为了得到金钱，你会把无菌注射器的针头扎入一个小孩儿的胳膊吗（关心与伤害）？别人送给你一台偷来的电视机，你会接受吗（公平与欺骗）？不论是保守派还是自由派，对于类似问题的回答都是否定的。*但对于下列问题，双方的答案出现了分歧：在国外的广播节目里，你是否会匿名批评自己的国家，将其称为“操控者”（忠诚与背叛）？出于喜剧表演需要，征得你父亲同意后，你会在他的脸上扇巴掌吗（权威与颠覆）？如果一部前卫派短剧要求演员像动物一样裸身爬行，像猩猩一样咕哝，你会参加表演吗（圣洁与堕落）？与自由派相比，保守派对这些问题更倾向于做出否定的回答（“当然不会”）。这是为什么呢？

之前提过，海特对此的解释是，自由主义者的道德味觉十分迟钝，他们的道德味蕾几乎有一半都失去了作用。为什么会这样呢？海特认为，西方道德哲学家以及其他受启蒙运动思想影响的人才是罪魁祸首。有孤独症倾向但又绝顶聪明的人认为避免伤害和公平不倚就是唯一重要的事情，以边沁和康德为典型代表。这种观点逐渐为多数人所接受。不久之后，又形成了一种新的文化思维：WeirD现代自由主义。（WeirD为西方的、有教养的、工业化的、富有的、民主的5个单词的首字母缩写。）根据海特的理论，如果要求社会保守派与自由派相互预测彼此对于道德问题所做出的回答，6种道德感受器全部正常工作的社会保守派将会略胜一筹。

那么我们应当如何看待自由主义者身上不完整的道德味觉呢？这是他们身上需要改正的缺点吗？从某种程度上来讲，是的。如果你是一名自由主义社会科学家，那么迟钝的道德味蕾便是你的一大劣势。如果你认为道德仅仅只是避免伤害、公平不倚，那么你可能会忽视很多人类行为，甚至对行为造成误读。同样的，如果你是一名努力争取中间选民的政治活动家，如果你的竞选广告中只包含了一种道德味觉感受器，而对手的广告中则包含了全部6种，那么你很可能会失去选票。此外，如前文所说，如果你是一位想要理解保守主义思想的自由主义者，对他们的道德品位有所了解也是非常有用的。但最为重要的问题依然没能得到解答：自由主义者是道德上有缺陷的人吗？我想答案是否定的，而且事实恰恰与此相反。

对于现代道德发展史，我有着完全不同的看法。就像新草地的寓言中所描述的一样，当今世界上汇聚了各种不同的部落，拥有不同的价值观与传统。启蒙运动中伟大的哲学家身处迅速缩小的世界当中，当时的人们被迫怀疑，他们自身的法律、传统，以及神祇是否真的优于其他部落。在那时，受到过良好教育的阶级刚刚兴起，技术（比如船只）与随之而提高的

经济生产力（比如全球贸易）将财富和权力赋予他们，他们也因而开始对传统的王权与宗教权力提出了质疑。最终，自然科学揭示了大自然的普遍规律，颠覆了古老的宗教教义，使人们能够从世俗的角度理解世界。哲学家们开始考虑，世间是否存在普遍的道德规律，就像牛顿的万有引力定律一样，适用于不同部落的一切成员，不论他们是否能够意识到，都不得不在这种规律的影响下生活。因此，启蒙运动中的哲学家所选择的道德味蕾并非一时兴起。他们出于高尚的动机，希望寻找到更深层次的普遍道德真理。他们所寻找的道德真理不受任何宗教教义的限制，超越了世间任何国王的意志。他们所寻找的便是我们所说的元道德：能够反映新草地上生活规律的一种泛部落哲学，或是后部落哲学。

和海特一样，有人可能会说，自由主义者的味觉不完整。但对于道德基础这一问题来说，少也可能是一种优势。与其说自由主义者的道德味觉不完整，不如说他们的道德味觉是更加灵敏的。

大多数情况下，美国的社会保守主义属于一个特殊的群体：欧美白人基督教部落。不幸的是，这个部落拥有很强的部族意识。当自然科学知识与部族教义发生冲突时，自然科学知识会被断然抛弃。此外，这个部落的成员将自身视为"真正的"美国人（即使不是明确表态，也有暗示表明），并将敢于挑战其部落信仰的人们视作外部入侵者。海特认为，美国的社会保守派对权威有着更多的尊重。从某种程度来讲，他的观点是正确的。社会保守派更加不愿掌掴自己的父亲，即使是开玩笑也不行，类似的行为都会引起他们的不适。但社会保守派并非不加选择地对所有权威表示尊重，他们仅会对自身部落认可的权威表示尊重（包括基督教的上帝、各种宗教领袖、政治领袖、父母等）。巴拉克·侯赛因·奥巴马作为一名土生土长的美国人，作为一名合法的总统，并没有得到美国社会保守派特别的敬意，反而常常受到他们的挑战。这类阴谋论本应是极右翼分子的做法，但2011年哥伦比亚广播公司与《纽约时报》共同进行的民意测验表明，共和党人中有45%都认为奥巴马所描述的身世可能是假的。同样的，与民主党人和无党派人士相比，共和党人对美国的国家权威几乎毫无敬意。大部分共和党人认为，尽管美国的穆斯林在政府中占有权威的一席之地，但他们是绝对不值得信任的。也就是说，社会保守派对权威的尊敬与他们对神圣的执念一样，都深植于部族主义的情感之中。（如果你认为先知穆罕默德是神圣的，那么你便不能够执掌政权。）最终，也是最为明显的一点，美国社会保守派对忠诚的关注也仅限于部落内部。他们并不认为每个人都应当忠于自己的国家。比如说，如果伊朗人想要对政府提出抗议，他们便会表示支持。

总之，美国社会保守派并没有为权威、圣洁和忠诚赋予特殊含义，他们不过是效忠于部落利益的群体。他们对部落内部的权威、信仰以及部落本身

保持忠诚。他们并非邪恶，但却是狭隘的，忠于部落的。从这个角度来讲，他们与世界上其他的社会保守派并无两样，包括阿富汗的塔利班组织与欧洲的民族主义者。海特认为，自由派应当更加主动地对社会保守派妥协。我对此持不同意见。短期来看，妥协也许是必要的手段；但长远来看，我们的策略应当是说服部落内的道德家减轻部族主义倾向，而并非向他们妥协。

我认为部族主义在群体层面的核心是利己主义，我不相信利己主义能够成就更大范围的利益，因此我没有成为社会保守派的成员。事实也有力地支持了我的观点。在丹麦、挪威、瑞典等完全的自由主义国家，只有很少一部分人宣称自己信仰上帝。如果自由派对美国社会的道德观造成了威胁，那么这些国家岂不是应当坠入地狱。但事实恰好相反，这些国家的犯罪率极低，学生的学业成就极高，居民生活水平和幸福感也都位居世界前列。海特认为，美国政治环境中，需要以保守派来制衡自由派，就像用“阳”来制衡“阴”一样。若真是如此，那么斯堪的纳维亚半岛是否也应当营造同样的状况？政治上没有达到平衡的丹麦是否应当从美国乡村进口基督教基要主义，用以平衡政治？在“剑桥人民共和国”中，担任公职的没有一位是共和党人，但在三大信用评级机构对美国各城市做出的评估中，剑桥便是为数不多获得3a评级的城市之一。

当然，社会保守派有很多优点值得自由派学习。正如海特所说，社会保守派善于和睦相处。与典型的自由派相比，保守派是好邻居，更加愿意为社区奉献出自己的时间和金钱。他们知道如何积累社会资本，如何建立社交网络和社会机构，如何建立信任、展开合作。也就是说，社会保守派善于避免最初的公地悲剧。然而，避免常识道德悲剧这个现代悲剧却并非他们所长。作为一名自由主义者，我敬佩保守派在当地的教堂中投入的社会资本，也希望自由派能够建立起同样密集有效的社会网络。但对于教堂中宣扬的有关堕胎、同性恋以及世界如何产生的观点，我并未表示认可。

效忠于部落利益的人并不是身边唯一的保守派。崇尚个人主义的北方牧民已经登上世界舞台，形成了另外一个重要的元部落。他们是自由主义者，自由市场派，“古典自由主义学派”，在社会和经济问题上都希望尽可能地减少政府干预。他们想要降低税收，减少社会公共项目，精简规章制度，限制财富的重新分配。他们还想获得堕胎的权利、吸食大麻的权利以及自由婚姻的权利。自由意志论者（如我所描述的这样）是部族意识最为淡薄的一个群体，即使是对现代自由派采取的温和的集体主义，他们也是唯恐避之不及。

那么深度实用主义者为何不选择自由意志论的立场呢？事实上，深度实用

主义者应当站在自由意志论的立场上。人类在政治上所有的选择可以从不受限制的自由市场式资本主义一直过渡到共产主义，我所选择的自由派与今天的自由意志论者更加相似，与往昔的共产主义者（源自经不起推敲的权利）则差异更大。南方牧民所崇尚的成熟的集体主义已经死去。当下的问题并非选择是否支持自由市场式的资本主义，而是是否应当采用集体主义机构对自由市场式的资本主义进行调节，调节的力度究竟如何？例如，帮助穷人、免费公共教育、国家医疗保险、累进税制等问题。

对某些自由意志论者来说，他们的政治主张事关基本权利。他们认为，将某人辛苦劳动挣得的钱拿走，转而分配给别人是完全错误的做法。政府没有权利命令人民何事应做，何事不应做。基于前文的观点，我对此持反对态度。我们无法直截了当地证明何人拥有何种权利。在经济问题上，自由意志论观点以公平的世界为前提：如果政府对市场的干预是不公平的，那么市场本身必然是公平的。赢者得到应有的奖赏，输者失去应有的财物。但我并不相信世界是公平的。包括我在内的很多人，在生命伊始便占据了巨大的优势。有些人克服重重阻碍，终于获得成功，但这并不意味着我们可以忽略他面临的不利条件。罗恩·保罗认为，如果一个人太过愚蠢，拒绝购买医疗保险，那么政府便不必对他负责。但这个人的孩子怎么办？那些出身贫寒，家长无力购买医疗保险的孩子们怎么办？政府应当任由他们死去吗？这些自由派观点已是老生常谈，我不再赘述。除非你相信世界是公平的，或者你认为只要稍加努力，一切社会经济问题的不平等就能被轻而易举地扫平。若非如此，那么基要主义者从权利的角度为自由意志论政策所做出的辩护便是无本之木。

功利主义者认为，自由意志论政策能够成就更大范围的利益。北方牧民认为，长远来看，对聪明勤劳的人施以惩罚、对愚蠢懒惰的人予以奖赏，这种做法对任何人都是不利的。罗恩·保罗认为，有些人会做出愚蠢的决定，这并不是一件好事。但倘若一个社会承诺帮助那些不愿自助的人，那么这个社会终将走向毁灭。一名社会保守派抗议者在条幅上写道：“分享工作精神，拒绝分享财富”。

有些情况下，我相信自由意志论者是正确的，至少比很多自由派的观点更加正确。在公立学校中引入竞争听上去是个好主意。从基本原则出发，我并不反对建立合法的人类器官市场。但我担心的是，由此引发的器官不当采集以及相关的暴力事件造成的负面影响过大，超过了人们更加便捷地获得器官所带来的好处。尽管可能面临自由派的抗议，我还是希望卖淫能够获得合法地位，并有相应的规章制度进行约束。抵制国外血汗工厂的产品对贫穷国家的工人可能会有害无利。将欧元作为欧洲的统一货币也许是勇敢而睿智的一个进步，也可能是通向超级集体主义这条错误道路的一次探险。时间自会证明一切。个人主义与集体主义之间的理想界限究竟在哪

里？我并不知道，也不想假装知道。但我对道德心理学略有了解，据此做出的判断多数会向左翼倾斜。人们想象中的功利主义是反对“大政府”的，我怀疑人们提出的论据事实上只是合理化的理由。我并不否认自由派也会为自己的观点寻找合理化解释（如前所述），我也不否认有诚实自觉的人对小政府表示支持。我想要表达的观点或假设是，很多反对政府的情绪与其初衷并不相符。类似的情绪大致可以分为两类。

首先，社会保守派为何反对“大政府”？显然，社会保守派与自由意志论者不同，不会坚定地支持个人主义。我猜测，社会保守派对美国联邦政府保持警惕的原因与他们对联合国保持警惕的原因相容。两者都是跨部落的权力机构，有意愿也有权利将取自“我们”的东西分给“他们”。（或者将“他们”的价值观强加给“我们”。）将金钱捐给当地教堂或者其他机构，为同部落的族人服务，社会保守派对此毫无意见。但联邦政府从他们口袋里拿走的钱并没有用于帮助辛苦工作的人，而是落到了骗取福利的“福利女王”手中，落到了“他们”的手中。我想，美国东部曾实行奴隶制的几个州成为共和党的票仓绝非偶然。对大多数人来说，哲学层面上对“大政府”的反对依据大多来源于部族主义。联邦医疗保险等能使“我们”获得明显、直接利益的政府项目不仅不会被阻碍，而且还会得到社会保守派的细心呵护。（一位愤怒的保守派曾在市政大会上说道：“将政府的手从我的联邦医疗保险上拿开！”）

另一部分坚决反对“大政府”的群体则是有钱人。他们希望降低税收，精简规章制度，尽可能减少社会项目。他们便是那众所周知的1%：人数不多，但重权在握。沃伦·巴菲特有句著名的话代表了其余99%的美国人的观点：如果亿万富翁纳税的税率低于他的秘书，那么一定是有什么地方出了问题。然而，如果你认为世界上最为富有的人应当因自己的智慧与勤劳获得额外奖赏，那么类似的政策对你来说便不足为奇。在我看来，这种政策太过自私，在自由派的主张中，我的观点绝非前无古人、后无来者。但我也认为，这种政策的倡议者是真诚的。众所周知，米特·罗姆尼（Mitt romney）曾将47%的美国人称为“不负责任的吃白食的人”，满屋的有钱人对这种观点都十分满意。但人们常常忽略的是，坐在罗姆尼屋子里的都是大选赞助者。米特·罗姆尼最为钟爱的听众并不会直接做出自私的行为。即使是精神失常的人也不会在竞选晚宴上一举掷出5万美元。这笔钱花在别的地方能让人获得更加可靠的收益*，或者享受更加有趣的时光。我相信，米特·罗姆尼和他的富人朋友们发自内心地相信，他们的行为能够成就更大范围的利益。这不是简单的利己主义，而是有倾向性的公平。

有些人年薪高达300万美元，典型的美国工人年薪仅为3万美元。这就是自由市场。我愿意相信的是，总体来说，年薪上百万的人比普通工人工作更加努力，他们应当得到更多的回报。但我不相信他们工作的努力程度能够

达到普通工人的100倍。我也不相信超级富豪们一周的工作量比普通工人一年的工作量还要多。富人们也许确实应得更多的财产，但他们也无疑从好运当中受益良多。世界上很多公立学校缺乏资金，无法给老师们发放与其工作相匹配的合理薪水；数十亿的孩子没有犯任何错误，但却生来贫困。面对这些，世界上最幸运的一部分人有什么理由将好运气独自享用呢？让富人拿出一些金钱对他们几乎没有损失；但是以合理的方式为穷人提供资源与机遇却会让他们受益匪浅。这不是社会主义，而是深度实用主义。

首先，我会将自己对于政治心理学的理解与乔纳森·海特的理解进行比较。从前文的描述中看，你可能会认为海特是一名坚定的保守派。但事实并非如此。他属于中间派，有时则会成为矛盾的自由派，归根结底，他支持一切符合功利主义思维的事件。*海特从根本上对功利主义的支持看似自相矛盾，但却具有一定的启发性。

海特认为，自由派的味觉并不完整，而功利主义的味觉则是最不完整的。有人在杰里米·边沁身故之后，将其诊断为阿斯伯格综合征患者，认为边沁患有轻度自闭症，与社会环境脱节。海特引用了这项研究，认为边沁的精神状况影响到了他的哲学思想，导致他将关乎道德的一切问题压缩为单一的价值观。海特用厨房作为比喻，设想出一桌“功利主义宴席”。边沁的厨房只能用一种十分蹩脚的哲学思维，激活一种道德味觉感受器，就像只会做甜食的饭店一样。但后来，海特又在自己的书中写道：

在个人主义者的私人生活中，我不知道哪一种标准道德理论是最好的。但如果我们在西方民主国家谈起制定法律和落实公共政策，同时想要保持一定的种族与道德多样性时，我认为功利主义是目前最好的选择。

这是怎么回事？在“我们应当怎么做？”这个终极问题面前，患有自闭症的哲学家提出的理论似乎又成了一贯正确的观点。在我看来，之所以会发生这种情况，是因为海特使用了大脑中不同的道德罗盘。

现代牧民的道德感觉十分强烈，有时这种感觉还会大相径庭。不幸的是，我们无法各行其是。那么该怎么办呢？海特认为，第一步是增进相互理解，认识到我们来自不同的道德部落，每个部落都有各自的想法，但这还不够。我们需要一种共同的道德标准，也就是元道德，从而帮助我们和睦相处。将幸福感最大化这个目标并不是对单一道德价值的盲目赞颂，也绝非用一个部落的价值观压制其他部落。这是使用通用货币进行衡量的结果。这是一种价值标准，我们能够以此为标准衡量其他价值观，不再盲目妥协，而是进行有原则的妥协。海特曾说过，“人性中有90%是黑猩猩，

10%是蜜蜂”。他的意思是，人类大都是自私的，但我们的本性中也有一部分部族精神，保卫着我们各自的蜂巢。我认为海特对人性的阐释并不完整。我们大脑中哪一部分支持将全部人类的幸福感最大化？我们既不是黑猩猩也不是蜜蜂。这种理想中的元道德完全是人类自己创造出来的，是抽象推理的产物。如果我们被利己主义和部落直觉所限制，人类便会止步不前。但幸运的是，不论我们是否情愿，人人都有能力切换到大脑的手动模式进行思考。

短期来看，道德推理的效率十分低下。*我想海特之所以会低估其重要性，便是由于这个原因。如果一位牧民心中已经认定某件事是正确或者错误的，即使你提出的论点再好，在当时当日改变他的想法也是十分困难的。但好的论点就像和风细雨，年复一年地荡涤着大地，事物的发展趋势最终也一定会被改变。**这个过程中的第一步，是对本部落的信仰提出质疑的意愿。在这一点上，有一点自闭的倾向或许是件好事。1785年前后，同性恋可被判处死刑，边沁写下了这样一段话：

当今欧洲各国对同性恋所实施的惩罚十分严厉。为了给这种惩罚找到足够的依据，我已经苦思冥想很多年。但根据功用的原则，我没能找到任何依据。

手动模式的道德思维要求人们勇敢且有毅力。密尔在1869年出版了为妇女争取权利的经典著作《论妇女的从属地位》（*The Subjection of Women*），其妻子哈莉特·泰勒·密尔可能也参与了写作过程。在引言部分，密尔写道：

倘若有人认为我的信念因为没有充分的论据支撑，或是因为论据不够明确，而面临重重阻碍，那么他一定是错误的。一旦人类的多种情感面临挑战，阻碍便产生了，对于任何事情都是如此。只要情感还存在，它便会不断抛出新的论据，修建新的防御工事，对原有的观点修修补补。

当今社会，我们中间确实有些人坚决维护同性恋和女性的权利。但我们诉诸情感处理这件事之前，在我们的情感认可他们的“权利”之前，总需要有人进行思考。我是一名深度实用主义者，也是一名自由主义者。因为我相信这条进步之路，在这条路上，我们任重道远。

# 第12章 傻瓜道德模式之外：6条规则

最初，空间中只存在宇宙原始汤。相互合作的分子结合成为大分子，有些大分子能够自我复制，并在外围形成一层保护膜。相互合作的细胞融合成为复杂细胞，然后相互合作形成细胞团。生命开始向复杂形式进化，一个又一个神奇角落出现，使个体的牺牲能够换来集体的成功，从蜜蜂到倭黑猩猩，无一例外。然而根据生物进化规则，相互合作的有机体并不会进行普遍合作。合作以竞争武器的身份产生并发展，是一种战胜他人的策略。因此，合作到达最高层面时，便不可避免地遭遇瓶颈。“我们”与“他们”之间产生分歧时，偏爱“我们”的力量便会阻碍合作。

有些动物进化出了大脑，作为计算控制中心，接收信息并据此指导行为。多数动物的大脑都采用反射型思维，能够根据接收到的信息自动产出结果，但却无法对事件进行反思，也无法创造新的行为。但人类的大脑则进化出一种完全不同的智慧。人类能够进行多方面的推理，解决反射型思维无法解决的复杂新奇的问题。快速反应与细思慢想的组合是制胜的关键，但同时也是十分危险的组合。凭借大脑，人类打败了自然界中绝大多数的敌人。我们能够生产足够的食物，还能够建造住处，保护自己。我们胜过了绝大多数天敌，从狮子到细菌。今天，站在我们面前的最强大的敌人便是我们自己。几乎所有的重大问题都因人类的选择而起，或是可以通过人类的选择而避免。

近些年来，人类已经在很大程度上减轻了彼此的憎恨，温和的商贸往来取代了战争；民主取代了独裁；科学取代了迷信。但我们还有进步的空间。我们面临贫困、疾病、战争、剥削、个人暴力等经年不息的全球性问题；还有气候变化、使用大规模杀伤性武器的恐怖行为等迫在眉睫的全球性问题；以及生物伦理学、大政府与小政府的选择、宗教在公共生活中的角色等现代生活所独有的道德问题。何种选择能让我们进步？

人类的大脑与其他器官一样，都是为了传播人类基因而存在。出于众所周知的原因，大脑使人具有自私的冲动，自动模式下的思维促使人为了生存繁衍而努力。基于一些不太明显的原因，大脑也会促使我们关心他人，关注他人的行为是否与自己相同。我们拥有同理心、爱情、友谊、愤怒、社会厌恶、感激、复仇、荣誉感、内疚、忠诚、谦逊、敬畏、论断是非、说短道长、窘迫，以及义愤等情感。这些人类心理中的共性使“我们”胜过了“我”，引导人类进入神奇角落，避免了公地悲剧。

所有健康人的大脑都拥有这些认知机制，但其工作方式却各不相同。不同的部落拥有不同的合作条件。关于人情世故，关于如何有尊严地面对威

胁，我们的观念与感受也各不相同。我们拥有不同的“专有名词”，不同的地方道德权威。我们都有部族主义情结，优先考虑“我们”而不是“他们”的利益，这是本性使然。即使我们认为自己公平不倚，但潜意识里，我们所谓的公平也是与“我们”的概念更加一致的公平。因此，不同的道德部落无法在是非判断上达成一致，我们面临着常识道德悲剧。

解决问题的关键往往在于正确地构建问题。在本书当中，我试着为人类最大的道德问题建立起一个思维框架。如前所述，人类面临的两大根本道德问题是“我”与“我们”之间的矛盾（公地悲剧）以及“我们”与“他们”之间的矛盾（常识道德悲剧）。人类大脑也包括两种本质不同的道德思维：快速反应（自动模式的情感）与细思慢想（手动模式的推理）。一切的关键都在于选择正确的思维模式解决相应的问题。面对“我”与“我们”之间的矛盾，请快速反应；面对“我们”与“他们”之间的矛盾，请细思慢想。

现代牧民需要放慢速度，努力思考，但我们首先需要找到正确的方法。如果使用手动模式思维来描述我们的道德感觉，或者为其寻找合理化解释，结果仍将是一无所获。我们不应试图将自动模式思维的成果进行排列或为其寻找理由，我们应当冲破限制。如此看来，问题的答案似乎已经十分清晰：我们应当将各自不同的部落情感搁置一边，选择整体上能够产生最好结果的行为。但问题在于：什么才是最好的？

我们所看重的一切几乎都是重要的，因为它们都会对我们的经历产生影响。因此，我们可以说，在所有人生活质量同等重要的前提下，能够使我们享受最好经历的便是最好的。边沁和密尔将这个绝妙的想法发展成为一种系统的哲学思想，并为之起了一个糟糕的名字。从那以后，人们对这种思想便一直存有误解，低估了它的价值。然而，这一思想的问题远不止糟糕的市场营销手段。人类的直觉本身并不是为了建立完善的道德哲学而存在。因此，真正完善的哲学思想必定会触动我们的神经。这种情况有时发生在现实世界，但更多的问题则出现在哲学思维试验中，我们会人为设置场景，将人性中最强烈的情感与更大范围的利益相比较。我们之所以会低估功利主义，是因为我们过分高估了自己的内心思维。我们错误地将直觉视为通向道德真理路上可靠的向导。正如契诃夫所说，我们需要了解自己，才能变得更好。

这本书很长，也很复杂，但我从一开始便做出承诺，要明晰地向你呈现。读到这里，我希望你对道德问题已经有了更加明晰的看法。我希望你能够看到常识道德悲剧在身边逐渐展开，认识到导致这种悲剧的根本原因在于人类情感，只有逻辑思维才能推动我们继续向前。我们已经介绍了很多观点，提到了很多“主义”。作为一名社会科学家兼实用主义者，我深知理论

与实践之间的鸿沟。我们的理想若要产生深远影响，不仅需要通过我们的“主义”阐释清楚，更要通过习惯深入人心。因此，在本书的最后，我为新草地上的生活提出了一些简单实用的建议。

## 给现代牧民的6条规则

### 规则1：面对道德争议，可以向道德直觉问策，但不要完全照办*

人类的道德直觉是绝妙的认知机制，经过了上百万年的生物进化、上千年的文化进化以及多年个人经验的积累。在个人生活中，你应当信任道德直觉，对手动模式思维心存警惕，因为手动模式思维总会寻找办法将“我”的利益置于“我们”之上。但面对道德争议，面对“我们”与“他们”之间的矛盾，便应当求助于手动模式。如果情感道德罗盘指向了不同的方向，那么总有一方是错误的。

### 规则2：权利不是好的论据，但却能终止辩论

我们没有直截了当的证据证明何人应当拥有何种权利，也无法将权利进行比较。我们热爱权利（还有它老套的姐姐：义务），因为权利能够将我们的主观感受变为抽象的道德客体，是我们能找到的最便捷的合理化解释。不论这样的客体是否存在，争论都会变得毫无意义。我们可以将“权利”作为盾牌，保护我们已有的道德进步。理性的辩论过后，我们也可以将“权利”作为言论武器。但我们对“权利”的使用应当十分慎重。即使真的使用，我们也必须清楚自己的目的：当我们诉诸权利言论时，并不是在为进一步辩论提出论据，而是在宣布辩论的终止。

### 规则3：注重事实，对自己和他人提出同样的要求

对于深度实用主义者来说，如果不了解一件事的工作原理和可能产生的效果，便无法判断这件事情是好是坏。然而，生活中的大多数人都会对自己知之甚少的政策妄下断言，从环境法规到医疗保险体系。公共道德辩论中，这种情况可能更甚。我们应当对自己和他人提出要求，不仅知道我们支持或反对哪项政策，还要知道这些政策的工作原理。不论一件事有效或是无效，我们都应当拿出证据，或要求别人拿出证据进行证明。不论是理论层面还是现实层面，如果我们对某件事缺乏了解，那么便应当效法苏格拉底的智慧，承认自己的无知。

### 规则4：警惕带有偏见的公平

公平有很多表达方式。潜意识中，我们总会倾向于选择最适合自己的那个版本。因为带有偏见的公平也是一种公平，看穿偏见总是困难的，特别是反省自身的时候。作为个体，我们会这样做，作为各自部落的忠诚成员，我们也会这样做。有时候，我们甚至会做出牺牲，使所属部落能够达成更大程度上带有偏见的公平。我们所表现出来的，是一种带有偏见的无私。

规则**5**：使用通用货币

我们在权利和正义的问题上永远无法达成一致，但有两件更加基本的事情，将我们紧密相连。其一，人类经历的起起伏伏。所有人都想要幸福，不愿受苦。其二，我们都能够理解黄金法则，理解其背后隐含的公平的理想。将这两点结合起来，我们便得到了通用货币，可以据此做出有原则的妥协。我们便可以忽视部落直觉的反对意见，协商一致，选择最好的、让我们整体上最幸福的做法。

为了找到最好的做法，我们需要一种价值的通用货币，但我们也需要一种事实上的通用货币。人类知识浩如烟海，但人们至今为止最为信任的知识依然是科学，这个选择十分合理。但科学并非绝对正确，当科学与部落信仰发生冲突时，人们会立即否定科学知识。但不管怎样，几乎所有人都会选择适合自己的科学证据。（如果在明天，可靠的科学家宣称地球事实上只有几千年的历史，创始论者会不会高兴地跳起来呢？）其他任何一种知识都无法享有科学的特权。在各自的部落里，在我们的心中，我们可以选择相信任何事物。但在新草地上，显而易见的事实构成的通用货币才是真理的判断标准。

规则**6**：给予

作为个人主义者，我们并不是生活中规则的制定者。但我们每个人都会做出一些生死攸关的重大抉择。生活富足的我们只要做出一点牺牲，便能够显著改善他人的生活。作为部落生活中的一员，我们对远方“以数字代表的”陌生人并不会产生过多的同情心。但几乎没有人能够真诚地宣称，我们最为奢华的享受比救人一命更加重要，*为一位无力支付医疗保险和教育费用的人提供一个美好的未来也比奢华的享受更加重要。我们可以欺骗自己，声称自己的捐赠无关紧要。如果我们在哲学层面小有造诣，还可以为我们自私的选择找到合理化依据。但最为诚恳开明的做法，无过于承认我们的习惯，接受严酷的现实，然后尽力做出改变。因为不太成功的诚恳努力远胜于成功的否认。

伊曼努尔·康德曾为“头顶的星空”和“心中的道德律”发出感叹。这种情感十分可爱，但我却无法完全苟同。人类在很多方面都是无与伦比的，但心中的道德律却是一件祸福参半的事。对我来说，更加值得感叹的是人类对心中的定律提出质疑，并代之以更好定律的能力。自然世界中合作无处不在，从微小的细胞到集结的狼群。但不论是多么默契的合作，其最初目的都是为了在竞争中胜出，与道德并不相干。但人类凭借极其发达的灵长类大脑，以某种方式发现了自然规律背后的抽象规则，将其改造后为己所用。在这些草地上，一些新的事物正在阳光下蓬勃生长：一个全球部落正在寻找成员。他们不是为了在竞争中胜出，只是单纯地以善之名行事。

# 致谢

首先我要感谢我的父母，劳里·格林和乔纳森·格林，他们一直鼓励我独立思考。由于他们不辞辛劳、默默奉献，我才得以全力工作，实施自己的想法。对我的父母，对永远爱我、支持我的哥哥丹和姐姐利兹，我的心中充满感激之情。还要感谢我的嫂子兼好友萨拉·斯腾伯格·格林，感谢家庭的新成员亚伦·法尔楚克。

前行路上，许多优秀的导师和出色的同事曾给我激励，认真考虑我尚未成熟的观点，并教给我思考的方法。我对本科时代引我入门的导师乔纳森·伯龙、保罗·罗津、阿马蒂亚·森、艾利森·西蒙斯以及德里克·帕菲特永怀感激。还要感谢我的研究生导师戴维·刘易斯和吉尔伯特·哈曼。也要感谢彼得·辛格，他与我同在一个联合会，为我提出许多宝贵的建议，给予我暖心的鼓励。我要感谢普林斯顿大学的哲学社团，我在那些日子里参与了有益的争论，度过了美好的时光。我要向博士后（非正式博士）导师乔纳森·科恩表达诚挚的谢意，感谢他当初选择了我，带我领略科学的艺术，向我展示大脑与思维的奇妙关系。同时，也要感谢利·耐斯特龙、约翰·达利、苏珊·菲斯克等普林斯顿大学心理学系的老师们，感谢你们为我提供了第二个家，帮助我探寻自身的发展。从研究生阶段开始，道德心理研究小组便成为我的另一个家，小组中的成员既是哲学家也是科学家，这群快乐的人让我的眼界进一步扩宽，心灵得到激励。尤其是最近几年，道德心理研究小组中的“部落长者”史蒂芬·斯蒂奇、约翰·多丽丝、肖恩·尼科尔斯以及沃特·辛诺特·阿姆斯特朗都给予我莫大的指导与支持。最后还要感谢哈佛大学心理学系为我这样一位不成熟的哲学家提供机会，感谢各位教师的激励与指导，尽管我学识尚浅，他们依然鼓励我写下此书。马扎林·巴纳吉、乔希·巴克霍茨、兰迪·巴克纳、苏珊·凯里、丹·吉尔伯特、克里斯汀·胡克、史蒂芬·科斯林、温迪·门德斯、杰森·米歇尔、马特·诺克、史蒂芬·平克、吉姆·斯达纽斯、杰西·斯内德克、菲利克斯·瓦纳肯、丹·维纳等人曾给予我许多建议与支持，对此我表示由衷的感谢。还要感谢马科斯·巴泽曼，感谢系部办公室的行政人员，他们的工作非常出色。

感谢罗伯特·莎博尔斯基介绍我认识了布罗克曼股份有限公司的卡廷卡·马特森。卡廷卡后来成为我的代理人，一直在身边给我支持和鼓励。感谢他给我机会，教我认知书的世界，他的耐心与聪颖给予我信心和力量，使我坚信本书能够出版。

像传说中的“特修斯之船”一样，我对本书的章节条目进行了无数次修正。经过不断完善，本书现在的内容已经和最初的设想大不相同。我永远感激

我的同事，是他们投入了大量宝贵的时间和精力，帮我完成并修订本书。他们完整地阅读了书稿，向我提出宝贵意见，感谢乔纳森·伯龙、马科斯·巴泽曼、保罗·布鲁姆、托马索·布鲁尼、阿列克·查克洛夫、摩西·科恩·埃利亚、费里·卡什曼、约翰·多丽丝、丹·吉尔伯特、乔纳森·海特、布雷特·哈尔西、安德里亚·赫伯雷恩、安娜·詹金斯、艾利克斯·乔丹、理查德·乔伊斯、西蒙·凯勒、乔舒亚·诺布、维克多·库玛、肖恩·尼科尔斯、乔·帕克斯顿、史蒂芬·平克、罗伯特·莎博尔斯基、彼得·辛格、沃特·辛诺特·阿姆斯特朗、塔姆勒·萨默斯、丹·维纳以及利亚纳·杨。此外，还有很多人对本书多个章节的撰写提出了宝贵意见，其中包括丹·埃姆斯、科特·格雷、吉尔伯特·哈曼、丹·凯利、马特·科林斯沃斯、卡特里娜·科斯洛夫、林赛·鲍威尔、杰西·普林兹、托里·麦克吉尔、爱德华·麦谢利、罗恩·马龙、玛利亚·梅里特、艾利克斯·普拉基亚斯、艾丽卡·罗德、艾迪娜·罗斯基、蒂姆·施罗德、苏珊娜·西格尔、钱德拉·斯里帕达、史蒂芬·斯蒂奇以及瓦莱丽·提比略。（其中丹·吉尔伯特、史蒂芬·平克和彼得·辛格完整地阅读了本书的两部草稿，在此对他们表示由衷的感谢。）还有许多人曾对本书的撰写提供无私帮助，他们的名字没能出现在这里，我对此表示真诚的歉意。这本书是许多人集体智慧的结晶，每想至此，我都备感惶恐，只希望自己没有辜负众人的努力。每个转折时刻，我都会得到重要提醒，一次又一次地免于陷入窘境。

感谢企鹅出版社的才俊们对我的帮助，尼克·特劳维恩的初审意见使我备感鼓励，为我指明了方向；维尔·帕尔默严谨的修订使我的作品增色不少。感谢我的编辑埃蒙·多兰，虽然合作时间不长，但他提出的建议和投入的时间对我来说都十分宝贵。继特劳维恩和多兰之后，本杰明·普拉特也为本书提出了视角独特的见解，我们的合作富有成效且意义深远。本杰明为我指出了书中存在的不足与未来的希望，此外，他还意识到，我的观点其实始于书的中部。对他各个方面的洞察力，我深表钦佩。在企鹅出版社中，我最想感谢的是斯科特·莫耶，他为我这个默默无闻的博士后提供了著书立说的机会。过去的一年中，斯科特纡尊降贵，与其他人一同承担本书的编辑工作，在许多方面，他甚至比我更了解本书。对此我既深感惭愧，又受到激励，继续前行。感谢斯科特将其非凡的才华与我共享，感激他马梅特式的智慧与学识，也感激他自始至终对本书的认可。

过去的7年间，一批聪明且专注的年轻科学家作为研究生和博士后，同我一起奋战在道德认知实验室。我和他们自己都无法预知这本书会给他们带来多少负面影响。我至少能以本书作者的身份出现，但他们得到的却只有未完成的其他事情、疲惫呆滞的目光以及深深的歉意。他们在实验室工作时极其耐心，对我极其支持，我永远无法回报他们的善良。（我好像能听到他们说：“那不是真相！”）谢谢你们，安倍宣仁、埃莉诺·阿米特、里根·贝恩哈特、多纳尔·卡希尔、阿列克·查克洛夫、费里·卡什曼、史蒂芬·弗兰克兰、肖娜·戈登–麦基恩、萨拉·戈特利布、克里斯汀·马–克拉姆斯、乔·帕

克斯顿、戴维·兰德，以及阿米塔伊·沈哈夫。因为有你们，我每天都能快乐地开始工作。特别感谢负责脑部成像的史蒂芬·弗兰克兰，还要感谢萨拉·戈特利布不遗余力地为我们准备图表、笔记和参考文献。

我还要衷心地感谢我亲爱的朋友们，在本书的撰写过程中，他们的温暖和理解一直支撑着我。尼克尔·拉米、迈克尔·帕蒂、保拉·福斯、乔希·阿什利·巴克霍茨、布雷特·哈尔西、瓦莱丽·莱维特·哈尔西，谢谢你们。

我的孩子山姆和弗利达为我带来了无尽的欢乐，在他们降临人世之前，我从未体会过这样的欢乐，没有什么事比做他们的父亲更让人开心。为了这本书，我没能在他们的童年时光陪他们摘南瓜、在海边玩耍、讲睡前故事，这是我最为抱憾的事情。我想对他们说，从现在开始，爸爸会经常在家的。最后，我想用我所有的成就感谢妻子安德里亚·赫伯雷恩，她是我最好的朋友，也是我一生的挚爱，这本书是献给她的。安德里亚的智慧与才华使本书增色不少，没有她的付出，这本书也许根本不会成为可能。安德里亚负责打理我们的生活，丝毫不计较我为家庭带来的不便，使我能够安心写书。对于她为家庭付出的一切，我终生感激。

# 注释

vii **“man will become”:** Chekhov (1977) 27, quoted in Pinker (2002). Chekhov, A. (1977). *Portable Chekhov.* New York: Penguin. Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

## 前言 常识道德悲剧

6 **awash in misinformation:** The most notorious false claim is that Obamacare establishes “death panels” that decide who gets to live and die (FactCheck.org, August 14, 2009). Democrats have made some false claims, too; for example, about details concerning whether people can keep their health insurance plans under Obamacare (FactCheck.org, August 18, 2009).
7 **exchange with Texas congressman Ron Paul:** Politisite (September 13, 2011).
8 **enormous bets on housing prices:** Financial Crisis Inquiry Commission (2011).
8 **government bailed out several of the investment banks:** This move had bipartisan support but was favored more strongly by Democrats (US House of Representatives, 2008; US Senate, 2008).
8 **tippy top . . . 400 percent:** Krugman (November 24, 2011).
9 **“Occupy a Desk!”:** Kim (December 12, 2011).
9 **“class warfare”:** http://www.huffingtonpost.com/2011/10/06/herman-cain-occupy-wall-street_n_998092.html.
9 **the “47 percent”:** http://www.motherjones.com/politics/2012/09/watch-full-secret-video-private-romney-fundraiser.
9 **thanks to lower tax rates:** Buffett (August 14, 2011).
10 **“steal and rob people with a gun”:** ABC News (December 5, 2011).
10 **“a parasite who hates her host”:** *The Rush Limbaugh Show* (September 22, 2011).
10 **values may color our view of the facts:** Kahan, Wittlin, et al. (2011); Kahan, Hoffman, et al. (2012); Kahan, Jenkins-Smith, et al. (2012); Kahan, Peters, et al. (2012).
11 **“proper nouns”:** Strictly speaking, these are the *referents* of proper nouns.
13 **better at getting along:** Pinker (2011).
13 **modern market economies . . . of human kindness:** Henrich, Boyd, et al. (2001); Henrich, Ensminger, et al. (2010); Herrmann et al. (2008).
13 **twentieth century . . . approximately 230 million people:** Leitenberg (2003).
13 **conflict in Darfur . . . 300,000 people:** Degomme and Guha-Sapir (2010).
13 **A billion people . . . live in extreme poverty:** World Bank (February 29, 2012) reporting data from 2008.
13 **More than twenty million people are forced into labor:** International Labour Organization (2012).
13 **more calls from employers:** Bertrand and Mullainathan (2003).
14 **What are we doing right?:** Pinker (2011).
15 ***utilitarianism:*** John Stuart Mill’s utilitarianism and Charles Darwin’s theory of natural selection emerged around the same time and have had highly overlapping fan bases from the start, beginning with Darwin’s and Mill’s mutual admiration. This is not an accident, I think. Both groundbreaking ideas favor manual mode over automatic settings. For a nice discussion, see Wright (1994), chapter 16.

# 第一部分 道德问题

## 第1章 公地悲剧

19 **“The Tragedy of the Commons”:** Hardin (1968).

20 **central problem of social existence:** Von Neumann and Morgenstern (1944); Wright (2000); Nowak (2006).

20 **This principle has guided the evolution:** Margulis (1970); Wilson (2003); Nowak and Sigmund (2005).

22 **the larger party will kill:** Mitani, Watts, et al. (2010).

22 **a phenomenon known as cancer:** Michor, Iwasa, et al. (2004).

22 **Darwin himself was absorbed:** Darwin (1871/1981).

23 **“red in tooth and claw”:** A. L. Tennyson, *In Memoriam AHH,* in Tennyson and Edey (1938).

23 **“Morality is a set”:** This view originated with Darwin (1871/1981) and has become the consensus view among behavioral scientists in recent decades. See Axelrod and Hamilton (1981); Frank (1988); Wright (1994); Sober and Wilson (1999); Wilson (2003); Gintis et al. (2005); Joyce (2006); de Waal (2009); Haidt (2012).

24 **not assuming . . . group selection:** Even if morality evolved simply through individual selection, favoring capacities for reciprocal altruism, the same argument applies.

25 **Wittgenstein’s famous metaphor:** Wittgenstein (1922/1971).

25 **nature’s “intentions”:** Birth control might be used to enhance one’s long-term genetic prospects through judicious family planning, but it certainly doesn’t have to be used that way.

## 第2章 道德机制

28 **Prisoner's Dilemma:** Puzzles of the Prisoner's Dilemma form were devised by M. Flood and M. Dresher of the Rand Corporation. See Poundstone (1992).

31 **"Golden Rule" . . . every major religion:** Blackburn (2001), 101.

31 **kin selection:** Fisher (1930); Haldane (1932); Hamilton (1964); Smith (1964). This mainstay of evolutionary biology has once again become controversial. See Nowak, Tarnita, et al. (2010).

32 ***reciprocity*, or reciprocal altruism:** See Trivers (1971) and Axelrod and Hamilton (1981). Encounters between potential cooperators may be chosen or forced. See Rand, Arbesman, et al. (2011). Here, as in the above Art and Bud story, I've made their encounters chosen, to be consistent with the Prisoner's Dilemma story, but in standard models of reciprocal altruism, the encounters are forced by circumstances. In either case, the same reciprocal logic applies.

33 **variations on the Tit for Tat theme:** See, for example, Nowak and Sigmund (1993).

33 ***anger*, *disgust*, or *contempt*:** See Rozin, Lowery, et al. (1999) and Chapman, Kim, et al. (2009). Note that negative feelings such as anger and disgust are not perfectly interchangeable. Anger is an "approach" emotion, motivating active aggression. Disgust, in contrast, is a "withdrawal" emotion that originally evolved to expel contaminating substances, such as feces and rotten meat, from the body. Which of these negative attitudes is most strategically appropriate will depend on the relative costs and benefits of active aggression versus selective disengagement.

33 ***gratitude* . . . willingness to cooperate:** Rand, Dreber, et al. (2009).

33 **food-sharing behavior in chimpanzees:** De Waal (1989). See also Packer (1977) and Seyfarth and Cheney (1984).

34 **emotional dispositions that we inherited:** Gintis, Bowles, et al. (2005).

34 **things don't always go as planned:** Nowak and Sigmund (1992); Rand, Ohtsuki, et al. (2009); Fudenberg, Rand, et al. (2010).

34 **De Waal and Roosmalen:** De Waal and Roosmalen (1979).

35 **program might be called *friendship*:** Seyfarth and Cheney (2012).

35 **world of our ancestors . . . more violent:** Daly and Wilson (1988); Pinker (2011).

36 **"long pig":** Stevenson (1891/2009).

36 **birth of modern military training:** Grossman (1995).

36 **laboratory study . . . human aversion to violence:** Cushman, Gray, et al. (2012).

36 **Figure 2.2:** Adapted with permission from Cushman, Gray, et al. (2012).

37 **"lost" letters:** Milgram, Mann, et al. (1965).

37 **tips at restaurants:** Pinker (2002), 259.

37 **some researchers have questioned:** Cialdini et al. (1987).

37 **we *feel bad* for them:** Batson et al. (1981); Batson (1991).

37 **more likely to cooperate . . . in a prisoner's dilemma:** Batson and Moran (1999).

37 **one experiences the feelings of others:** In some cases, empathizing does not entail having the same feeling as the person with whom one is empathizing. For example, when one empathizes with a child who is scared, one need not be scared oneself.

37 **neural circuits that are engaged:** Singer, Seymour, et al. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303(5661), 1157–1162.
38 **Oxytocin . . . in maternal care:** Pedersen, Ascher, et al. (1982).
38 **Genes . . . oxytocin:** Rodrigues, Saslow, et al. (2009).
38 **more likely to initiate cooperation:** Kosfeld, Heinrichs, et al. (2005). But see Singer, Snozzi, et al. (2008).
38 **capacity to care about others:** De Waal (1997, 2009); Keltner (2009).
38 **Ladygina-Kohts:** As described by de Waal (2009).
38 **Arnhem Zoo:** de Waal (2009).
38 **without expectation of a reward:** Warneken et al. (2006, 2007, 2009).
39 **capuchin monkeys:** Lakshminarayanan and Santos (2008).
39 **empathy in rats:** Bartal, Decety, et al. (2011).
40 **MAD:** The formal theory behind MAD comes from von Neumann and Morgenstern (1944).
41 **emotional machinery that performs the same function:** Frank (1988). See also Schelling (1968). My example here follows Pinker (1997), who compares angry passions to the "Doomsday Device" in Stanley Kubrick's film *Dr. Strangelove.* The device is designed to automatically launch a nuclear counterstrike in the event of a first strike.
41 **not the only ones with a taste for vengeance:** Jensen, Call, et al. (2007).
41 **chimps do much the same in the wild:** De Waal and Luttrel (1988).
41 **breaking a promise:** Baumgartner, Fischbacher, et al. (2009).
42 **familial love and friendship . . . irrational:** Pinker (2008).
42 **in-house bank-robbing expert:** As in the case of Frank Abagnale. See Abagnale and Redding (2000).
42 **As Steven Pinker observes:** Pinker (2008).
43 **sometimes regard their leaders:** Henrich and Gil-White (2001).
43 **things larger than ourselves:** Keltner and Haidt (2003); Haidt (2012).
44 **Reputations can also enhance cooperation:** Nowak and Sigmund (1998).
44 **Kevin Haley and Daniel Fessler:** Haley and Fessler (2005).
44 **"Dictator Game":** Forsythe, Horowitz, et al. (1994).
45 **Figure 2.3:** Reprinted with permission from Haley and Fessler (2005).
45 **"honor box" for buying drinks:** Bateson, Nettle, et al. (2006).
45 **65 percent of their conversation time:** Dunbar (2004); Dunbar, Marriott, et al. (1997).
45 ***gossiping* . . . for social control:** Feinberg et al. (2012b); Nowak and Sigmund (1998, 2005); Milinski, Semmann, et al. (2002).
46 **transgressor appears to be embarrassed:** Semin and Manstead (1982); Keltner (2009).
46 **we're judgmental as *babies*:** Hamlin, Wynn, et al. (2007, 2011). See also Sloane, Baillargeon, et al. (2012).
47 **Figure 2.4:** Reprinted with permission from Hamlin, Wynn, et al. (2007).
47 **the *hindering* square:** In some versions of the experiment, the colors and shapes were reversed, showing that it's not just a preference for certain shapes or colors. In another version of the experiment, they showed that the infants prefer the helper to a neutral shape and prefer a neutral shape to a hinderer.
49 **enthnocentrism as universal:** Brown (1991).
50 ***parochial altruism*:** Bernhard, Fischbacher, et al. (2006); Choi and Bowles (2007).
50 **infants . . . language to distinguish:** Kinzler, Dupoux, et al. (2007). See also Mahajan and Wynn (2012).
51 **separating Us from Them:** McElreath, Boyd, and Richerson (2003).
51 **Implicit Association Test:** Greenwald, McGhee, et al. (1998); Greenwald and Banaji (1995).
51 **Try it yourself:** https://implicit.harvard.edu/implicit.
51 **whites have an implicit preference for whites:** Greenwald, McGhee, et al. (1998).
52 **children . . . same kind of race-based biases:** Baron and Banaji (2006).
52 **IAT developed for monkeys:** Mahajan, Martinez, et al. (2011).
52 **(Emily, Greg) . . . (Lakisha, Jamal):** Bertrand and Mullainathan (2003).
52 **stereotypically black facial features:** Eberhardt, Davies, et al. (2006).
52 **Google searches:** Stephens-Davidowitz (2012).
53 **sensitivity to race . . . group membership:** Kurzban, Tooby, et al. (2001).
53–54 **Classic studies by Henri Tajfel:** Tajfel (1970, 1982); Tajfel and Turner (1979).
54 **oxytocin . . . out-group members:** De Dreu, Greer, et al. (2010, 2011).
55 **"nasty, brutish, and short":** Hobbes (1651/1994).
55 **religion may be a device:** Wilson (2003); Roes and Raymond (2003); Norenzayan and Shariff (2008).
56 **wary of people who are not "God-fearing":** Gervais, Shariff, et al. (2011).
56 **enforced cooperation:** Boyd and Richerson (1992).
56 **pre-agricultural societies are rather egalitarian:** Boehm (2001).

56 **odds of getting punished . . . very high:** This assumes that the punished cannot or will not retaliate in full force against the punishers. If they can and do, cooperation may break down. See Dreber, Rand, et al. (2008); Hermann, Thöni, et al. (2008).

56 ***indirect reciprocity*:** Gintis (2000); Bowles, Gintis. et al. (2003); Gintis, Bowles, et al. (2005).

57 **"altruistic punishment":** Fehr and Gächter (2002); Boyd, Gintis, et al. (2003).

57 **people are . . . pro-social punishers:** Fehr and Gächter (2002); Marlowe, Berbesque, et al. (2008). But see Kurzban, DeScioli, et al. (2007).

57 **"Public Goods Game":** Dawes, McTavish, et al. (1977).

58 **contributions typically go up:** Boyd and Richerson (1992); Fehr and Gächter (1999).

58 **pro-social punishment is just a by-product:** Kurzban, DeScioli, et al. (2007).

58 **pro-social punishment evolved through:** Boyd, Gintis, et al. (2003).

59 **life on Earth . . . story of increasingly complex cooperation:** Margulis (1970); Nowak (2006); Wright (2000).

62 **familiar features of human nature:** Many of them appear on Donald Brown's (1991) list of human universals (e.g., "empathy," "gossip," "shame," "revenge"), and all of them are closely related to, if not logically entailed by, items on his list.

62 **feelings that do this thinking for us . . . series of studies:** Rand, Greene, et al. (2012).

63 **Figure 2.5:** All data reported in ibid. Decision time (in log10 seconds) is on the x-axes. Contribution level is on the y-axes, expressed as a percentage of the maximum (top) or probability of cooperation given a yes/no choice (all others). Top: A one-shot Public Goods Game. Middle left: First decision from a series of one-shot Prisoner's Dilemmas. Middle right: A repeated Prisoner's Dilemma with execution errors. Bottom left: A repeated Prisoner's Dilemma with or without costly punishment. Bottom right: A repeated Public Goods Game with or without reward and/or punishment. Dot size is proportional to number of observations, which are indicated next to each dot. Error bars indicate standard error of the mean.

64 **prohibition against eating cows may increase the food supply:** Harris, Bose, et al. (1966).

64 **blocking alternative routes to sexual gratification:** Davidson and Ekelund (1997).

65 **divine will and chance:** Dawkins (1986).

## 第3章 新草地上的争斗

69 **not about whether . . . but about *why*:** Pinker (2002).

69 **altruism within groups could not have evolved:** Choi and Bowles (2007); Bowles (2009).

70 **anthropologists studying small-scale societies:** Henrich, Boyd, et al. (2001); Henrich, Gil-White, et al. (2001); Henrich, McElreath, et al. (2006); Henrich, Ensminger, et al. (2010).

70 **"Ultimatum Game":** Güth, Schmittberger, et al. (1982).

71 **"Wall Street Game" or the "Community Game":** Liberman, Samuels, et al. (2004).

71 **Americans tend to give nothing:** List (2007).

72 **most cooperative . . . most willing to punish:** Henrich, McElreath, et al. (2006).

72 **Dictator Game . . . involves no *co*-operation:** The Dictator Game is about altruism, giving something for nothing. But from a mathematical, game-theoretical perspective, cooperation and altruism are (or can be) equivalent. This is because cooperation (the interesting kind) requires paying a personal cost to benefit others. To be cooperative is to be altruistic, and successful cooperation is just mutually beneficial altruism. (One notable difference between the Dictator Game and true cooperation games such as the Prisoner's Dilemma and the Public Goods Game, however, is that the size of the pie is typically not fixed in true cooperation games. That is, true cooperation games are not "zero sum.") One might say that cooperation is not altruistic if one expects to get something of equal or greater value in return for one's contribution, and that's fair. But one can say the same for altruism. In one-shot interactions, cooperation is altruistic, and in repeated interactions neither cooperation nor altruism is necessarily altruistic in the strongest sense.

72 **A more recent study:** Henrich, Ensminger et al. (2010).

73 **cooperation and punishment in a set of large-scale societies:** Herrmann, Thöni, et al. (2008).

73 **Figure 3.1:** Adapted with permission from ibid. (2008). Here I show data from only nine of the sixteen cities studied to more clearly illustrate three prominent trends.

74 **punish cooperators on the first round:** Ellingsen, Herrmann, et al. (2012).

75 **Figure 3.2:** Adapted with permission from Herrmann, Thöni, et al. (2008).

75 **Greece . . . had become financially insolvent:** BBC News (November 27, 2012).

76 **airing their prejudices:** There is a long history of scholars presenting research that is suspiciously consistent with their own ideological and cultural commitments. Take, for example, the case of Ellsworth Huntington, an eminent early-twentieth-century geographer, Yale professor, and president of the board of directors of the American Eugenics Society. (Yes, they had a society.) Huntington believed that economic development is

determined primarily by climate. More specifically, he concluded that New Haven, Connecticut (home to Yale University), has a more or less ideal climate for intellectual innovation and economic development.

And here I am, a secular Jew and professor at Harvard University, in metropolitan Boston, telling you that predominantly Muslim cities such as Riyadh and Muscat have cultures that thwart cooperation, while other cities have cultures that are highly conducive to cooperation—cities such as, oh, let's say, Boston.

Amid such suspicion, let me begin with three straightforward points about this research: First, I didn't conduct this research. Second, these results are not cherry-picked. As far as I know, there are no similar studies showing substantially different results, and if you want to see for yourself, you can hop on Google Scholar and search for them. Third, the results presented above are mixed. Riyadh and Muscat come out near the bottom in this study of cooperative behavior, but so does Athens, the birthplace of democracy and the cradle of Western philosophy. And while it's true that my hometown comes out near the top, so does Bonn, not far from the former sites of Nazi concentration camps. But beyond this particular set of studies and my relation to them, there's a more general question about how we ought to respond to scientific research that makes some people look in some ways better than others, particularly when the research's proponents are among those who come out looking better.

Let's start with the two extreme positions on this issue. At one extreme, we have complete deference: If someone with scientific credentials says that something is well supported by scientific evidence, then it must be true. For obvious reasons, this is not a good policy. At the other extreme, we have complete suspicion and skepticism: Anytime a scientist presents supposedly scientific research that makes some people look in some way better than others, we should be highly skeptical, and if the research makes its proponents look better than others, we should assume that it's just a bunch of politically motivated, self-serving, phony baloney.

As you might expect, I think that our attitude should be somewhere in the middle. To be open-minded, we must allow for the possibility that cross-cultural research will reveal cultural differences that make some cultures look better than others—not better in some ultimate sense, but better in some ways. Moreover, we must allow for the possibility that researchers who discover such differences may be among those who end up looking better. (For all I know, Huntington might be right.) At the same time, we should bear in mind that scientists are people, and that like all people they are subject to bias, including unconscious bias. (See later in this chapter.)

In reacting to cross-cultural social science, we should be clear about what follows from the science and what depends on our own moral assumptions. For example, the studies described above tell us that the people tested in Boston and Copenhagen made more money by cooperating in Public Goods Games than did the people tested in Riyadh and Athens. However, these studies don't tell us whether the people in Boston and Copenhagen played *better* or whether they have cultural attributes that are generally better to have. Those are value judgments that go beyond the data.

We should be aware that scientific studies may have serious flaws. (What if the people in Boston understood the directions better than the people in Athens?) At the same time, we should have some respect for the scientific process. Scientific journals select papers for publication based on anonymous peer review, and—believe me—scientists are plenty critical of one another, especially when they're anonymous. Papers published in respectable journals may have serious flaws, but they are unlikely to have *obvious* flaws. (As it happens, in all of the studies described above, participants were tested to ensure that they understood the directions, which is standard practice in this kind of research.)

In sum, bias in cross-cultural social scientific research is a legitimate concern, but the solution is not to dismiss all such research as biased political posturing. This is no better than blindly believing that everything well-credentialed scientists have ever said is true.

76 **studies examining cultural differences among Americans:** Cohen and Nisbett (1994); Nisbett and Cohen (1996).

78 **In surveys, southerners:** Cohen and Nisbett (1994).

78–79 **"Southern ideas of honor":** Fischer (1991).

79 **southern support for their anti-Soviet foreign policies:** Lind (1999).

79 **based on cooperative agriculture:** Nisbett, Peng, et al. (2001).

79 **American and Chinese . . . Magistrates and the Mob:** Described in Doris and Plakias (2007).

80 **"corrupt mind":** Anscombe (1958), cited in ibid. Note that Anscombe says this about killing an innocent person to quell the mob. It's not clear what she would say about imprisoning an innocent person.

80 **sensitive nature of these topics:** See "airing their prejudices" above in this chapter.

81 **reading of the Koran to non-Muslims:** Denmark TV2 (October 9, 2004) (translated with Google Translate).

82 **reward . . . to anyone who would behead "the Danish cartoonist":** *Indian Express* (February 18, 2006).

82 **boycotts of Danish goods cost . . . $170 million:** BBC News (September 9, 2006).

82 **YouTube video:** *International Business Times* (September 21, 2012).

83 **"If someone sues you":** *U.S. News & World Report* (January 30, 1995), cited by Bazerman and Moore (2006), 74.

83 **both versions of the question:** Hsee, Loewenstein, et al. (1999).

84 **experiments have documented this tendency:** Walster, Berscheid, et al. (1973); Messick and Sentis (1979).

84 **"Performance-based pay is fair":** Van Yperen, van den Bos, et al. (2005).

84 **series of negotiation experiments:** Babcock, Loewenstein, et al. (1995); Babcock, Wang, et al. (1996); Babcock and Loewenstein (1997).

85 **didn't know which side they would be on:** Note the parallel with Rawls's (1971) "veil of ignorance."

85 **better able to remember . . . material that supported their side:** Thompson and Loewenstein (1992).

86 **biased fairness . . . environmental commons problem:** Wade-Benzoni, Tenbrunsel, et al. (1996).

86 **negotiate over penalties for . . . criminals:** Harinck, De Dreu, et al. (2000).

89 **tribalistic brand of biased fairness:** Cohen (2003).

89 **It's all unconscious:** The results of Cohen's study are fascinating and sobering, but I would be amazed if they generalized to issues in which partisan disagreements are qualitative rather than quantitative. For example, I doubt that partisan repackaging could reverse people's views on abortion or gay marriage, at least not in the short run.

89 **believed that Saddam Hussein had been personally involved:** Millbank and Deanne (September 6, 2003).

90 **2008 World Public Opinion poll:** Reuters (September 10, 2008).

90 **students . . . watched footage from a football game:** Hastorf and Cantril (1954).

90 **more confident in their views . . . considering the mixed evidence:** Lord, Ross, et al. (1979).

90 **"hostile media effect":** Vallone, Ross, et al. (1985).

90 **people watched footage of protesters:** Kahan and Hoffman (2012).

91 **consensus among experts:** Intergovernmental Panel on Climate Change (2007); Powell (November 15, 2012).

91 **Republicans believe:** Jones (March 11, 2010), reporting on Gallup poll.

91 **commons problem of its own:** Kahan, Wittlin, et al. (2011).

94 **Figure 3.3:** Adapted with permission from Kahan, Peters, et al. (2012).

94–95 **In 1998, Republicans and Democrats . . . in 2010:** Dunlap (May 29, 2008) and Jones (March 11, 2010), reporting on Gallup polls.

95 **deep geologic isolation:** Kahan, Jenkins-Smith, et al. (2011).

95 **biased perception in the escalation of conflict:** Shergill, Bays, et al. (2003).

96 **Figure 3.4:** Adapted with permission from ibid. (2003).

96 **brain automatically anticipates:** Blakemore, Wolpert, et al. (1998).

97 **only partially aware of the contributions of others:** Forsyth and Schlenker (1977); Brawley (1984); Caruso, Epley, et al. (2006).

97 **As Steven Pinker explains:** Pinker (2011).

98 **some of our biggest problems:** Copenhagen Consensus Center (2012).

99 **"perfect moral storm":** Gardiner (2011).

100 **any .number of arrangements:** Singer (2004).

100 **the world's second-largest carbon emitter:** Union of Concerned Scientists (2008), citing data originally compiled by the US Energy Information Agency (2008).

100 **"cleaning up the world's air":** Second presidential debate in 2000, quoted in Singer (2004), 26.

101 **"An attempt to point out":** Fisher (1971), 113.

101 **"Officials think of themselves":** Ibid., 112.

101 **"Laying down the moral law":** Schlesinger (1971), 73.

# 第二部分 道德反应的快与慢

## 第4章 小火车的学问

106 **Jeremy Bentham and John Stuart Mill:** Mill (1861/1998); Bentham (1781/1996). The third of the great founding utilitarians is Henry Sidgwick (1907), whose exposition of utilitarianism is more thorough and precise than those of his more famous predecessors. Thanks to Katarzyna de Lazari-Radek and Peter Singer for highlighting many of the points on which Sidgwick anticipates points made here.

109 **without a girlfriend:** Or boyfriend, as the case may be.

111 **theory of parental investment:** Trivers (1972).

112 **in some birds and fish:** Eens, M., & Pinxten, R. (2000). Sex-role reversal in vertebrates: Behavioural and endocrinological accounts. *Behavioural processes*, 51(1), 135–147.

112 **Peter Singer first posed it:** Singer (1972).

113 **Two rivers, twenty rivers: It all sounds the same:** This attitude would make sense if the question were about making a small contribution to a larger (two-river or twenty-river) effort. But the question here is about how much you would pay to have the whole job completed if somehow your payment alone would do it.

113 **my first scientific publication:** Baron and Greene (1996).

113 **heuristic thinking:** Gilovich, Griffin, et al. (2002); Kahneman (2011).

113 **"The Trolley Problem":** Thomson (1985). The first papers on the Trolley Problem were by Philippa Foot (1967) and Thomson (1976). This gave rise to a large literature in ethics. See Fischer and Ravizza (1992); Unger (1996); and Kamm (1998, 2001, 2006).

114 **wrong to push the man off the footbridge:** Thomson (1985); Petrinovich, O'Neill, et al. (1993); Mikhail (2000, 2011); Greene, Somerville, et al. (2001).

115 **"Act so that you treat humanity":** Kant (1785/2002). This is the second of four formulations of the categorical imperative described in the *Groundwork*.

116 **people all over the world agree:** O'Neill and Petrinovich (1998); Hauser, Cushman, et al. (2007).

117 **the case of Phineas Gage:** See Damasio (1994) and Macmillan (2002).

118 **due to emotional deficits:** Saver and Damasio (1991); Bechara, Damasio, et al. (1994).

118 **"to know, but not to feel":** Damasio (1994), 45.

119 **Cognitive control:** Miller and Cohen (2001).

119 **color-naming Stroop task:** Stroop (1935).

120 **enabled by . . . DLPFC:** Miller and Cohen (2001).

122 **fMRI:** fMRI uses an MRI scanner of the sort routinely used in modern hospitals. For most clinical purposes, an MRI takes a still, three-dimensional picture of the body, a "structural scan." fMRI takes "movies" of the brain in action. The movies have a modest spatial resolution, composed of "voxels" (volumetric pixels) of about 2 to 5 millimeters. The temporal resolution is very low, with an image (a "frame" in the movie) acquired about once every 1 to 3 seconds. The images produced by fMRI look like pixilated blobs, which are typically overlaid on top of a higher-resolution structural scan, allowing one to see where the blobs are in the brain. The blobs are not the direct result of "looking" at the brain. They are the products of statistical processing. What a blob in a brain region typically means is that there is, on average, more "activity" in that region when someone is performing one task (e.g., looking at human faces) as compared with another task (e.g., looking at animal faces). The "activity" in question is the electrical activity of neurons in the brain, but this activity is not measured directly. Instead, it's measured indirectly, by tracking changes in the flow of oxygenated blood. For more information, see Huettel, Song, et al. (2004).

122 **including parts of the VMPFC:** Greene, Somerville, et al. (2001). Many other brain regions exhibited effects in this contrast, including most of what is now called the "default network" (Gusnard, Raichle, et al., 2001). Many of these regions appear to be involved not in emotional response per se, but in the representation of nonpresent realities (Buckner, Andrews-Hanna, et al., 2008).

122 **Our second experiment:** Greene, Nystrom, et al. (2004).

123 **We addressed this question in later work:** Greene, Cushman, et al. (2009).

123 **ice cream doesn't cause drowning:** I don't know who first used this example.

124 **follow-up studies of our own:** Some philosophers have raised doubts about the evidence for the dual-process theory (McGuire, Langdon, et al., 2009; Kahane and Shackel, 2010; Kahane, Wiech, et al., 2012; Berker, 2009; Kamm, 2009). For replies, see Paxton, Bruni, and Greene (under review); Greene (2009); and Greene (under review). For further details on Berker, see a set of notes (Greene, 2010) assembled for a conference at Arizona State University, available on my webpage or by request.

124 **patients with . . . (FTD):** Mendez, Anderson, et al. (2005).

125 **dilemmas . . . patients with VMPFC damage:** Koenigs, Young, et al. (2007); Ciaramelli, Muciolli, et al. (2007).

125 **sweaty palms:** Moretto, Ladavas, et al. (2010).

125 **same conclusion:** See also Schaich Borg, Hynes, et al. (2006); Conway and Gawronski (2012); Trémolière, Neys, et al. (2012).

125 **turning the trolley onto family members:** Thomas, Croft, et al. (2011).

125 **Low-anxiety psychopaths:** Koenigs, Kruepke, et al. (2012). See also Glenn, Raine, et al. (2009).

125 **alexithymia:** Koven (2011).

125 **physiological arousal . . . fewer utilitarian judgments:** Cushman, Gray, et al. (2012). See also Navarrete, McDonald, et al. (2012).

126 **gut feelings:** Bartels (2008).

126 **Inducing people to feel mirth:** Valdesolo and DeSteno (2006); Strohminger, Lewis, et al. (2011).

126 **the amygdala:** Adolphs (2003).

126 **psychopathic tendencies:** Glenn, Raine et al. (2009).

126 **amygdala . . . correlates negatively with utilitarian judgments:** Shenhav and Greene (in prep.).

126 **citalopram . . . fewer utilitarian judgments:** Crockett, Clark, et al. (2010).

126 **lorazepam has the opposite effect:** Perkins, Leonard, et al. (2012).

126 **role of visual imagery:** Amit and Greene (2012).

126 **"Do whatever will produce the most good":** By this I don't mean that people who make utilitarian judgments are card-carrying utilitarians, subscribing to the full philosophy. I mean only that they are applying an impartial "cost-benefit" decision rule.

127 **DLPFC . . . success in the Stroop task:** MacDonald, Cohen, et al. (2000).

127 **brain imaging studies . . . similar results:** Shenhav and Greene (2010); Sarlo, Lotto, et al. (2012); Shenhav and Greene (in prep.).

127 **simultaneous secondary task:** Greene, Morelli, et al. (2008). See also Trémolière, Neys, et al. (2012).

127 **removing time pressure and encouraging deliberation:** Suter and Hertwig (2011).

127 **experience of being led astray:** Method follows Pinillos, Smith, et al. (2011).

127 **tricky math problems:** Frederick (2005).

127 **people who solved these tricky math problems:** Paxton, Ungar, and Greene (2011). In the case of the *footbridge* dilemma, the tricky math problems didn't change people's judgments. Instead, we found that people who were generally better at solving the tricky math problems gave more utilitarian judgments in response to the *footbridge* case. See also Hardman (2008). Paxton and I also used the CRT method with a "white lie dilemma" devised by Kahane, Wiech, et al. (2012), a case in which the non-utilitarian response was alleged to be counterintuitive. Our results indicate the opposite, consistent with the original dual-process theory. See Paxton, Bruni, and Greene (under review).

127 **people who generally favor effortful thinking:** Bartels (2008); Moore, Clark, et al. (2008).

128 **moral reasons of which people are conscious:** Cushman, Young, et al. (2006).

128 **justifying that judgment in a consistent way:** Ibid.; Hauser, Cushman, et al. (2007).

128 **on a molecular level:** Crockett, Clark, et al. (2010); Perkins, Leonard, et al. (2010); Marsh, Crowe, et al. (2011); De Dreu, Greer, et al. (2011).

129 **moral judgments of medical doctors:** Manuscript in preparation, based on Ransohoff (2011). For a review of bioethical issues from a neuroscientific perspective, see Gazzaniga (2006).

129 **minimize the risk of actively harming:** It's often said that the Hippocratic Oath, taken up by doctors upon entering the profession, commands them to "First, do no harm." However, these words don't actually appear in the oath. See: http://www.nlm.nih.gov/hmd/greek/greek_oath.html.

## 第5章 效率、灵活与大脑的双加工机制

132 ***Everything Bug***: Winner (2004).

132 **one of the most important ideas to emerge:** Posner and Snyder (1975); Shiffrin and Schneider (1977); Sloman (1996); Loewenstein (1996); Chaiken and Trope (1999); Metcalfe and Mischel (1999); Lieberman, Gaunt, et al. (2002); Stanovich and West (2000); Kahneman (2003, 2011).

134 ***Thinking, Fast and Slow***: Kahneman (2011).

134 **get rid of the concept of "emotion":** Griffiths (1997).

135 **but it's not emotional:** Such processing may trigger emotional responses, but the visual processing itself is not emotional.

135 ***action tendencies***: Darwin (1872/2002); Frijda (1987); Plutchik (1980).

135 **Fear . . . enhancing the sense of smell:** Susskind, Lee, et al. (2008).

135 **influenced . . . decisions by influencing their moods:** Lerner, Small, et al. (2004).

137 **"slave of the passions":** Hume (1739/1978).

137 **cannot produce good decisions without . . . emotional input:** VMPFC patients like Phineas Gage are, in general, very bad decision makers. They can give reasons for choosing one thing over another, and the reasons they give often sound good. But these reasons are fragmented. Instead of adding up to a good decision, they float free in a jumble, resulting in foolish behavior. (See Damasio, 1994.) In a telling experiment, Lesley Fellows and Martha Farah (2007) showed that VMPFC patients are more likely than others to exhibit "intransitive" preferences—that is, to say that they prefer A to B, B to C, and C to A. With respect to decision making, this is the hallmark of irrationality. What's more, the DLPFC, the seat of abstract reasoning, is deeply interconnected with the dopamine system, which is responsible for placing values on objects and actions (Rangel, Camerer, et al., 2008; Padoa-Schioppa, 2011). From a neural and evolutionary perspective, our reasoning systems are not independent logic machines. They are outgrowths of more primitive mammalian systems for selecting rewarding behaviors—cognitive prostheses for enterprising mammals. In other words, Hume seems to have gotten it right.

137 **fruit salad or chocolate cake:** Shiv and Fedorikhin (1999).

138 **two different kinds of decisions:** McClure, Laibson, et al. (2004).

139 **yielding immediate rewards:** Here the immediate rewards are not so immediate. In a later study, using food rewards (McClure, Ericson, et al., 2007), they were more immediate.

139 **Figure 5.1:** Images adapted with permission from Ochsner, Bunge, et al. (2002); McClure, Laibson, et al. (2004); and Cunningham, Johnson, et al. (2004).

140 ***intra*personal . . . *inter*personal:** Nagel (1979).

140 **we see the same pattern:** Cohen (2005).

140 **reinterpret the pictures in a more positive way:** Ochsner, Bunge, et al. (2002).

140 **presented white people with pictures:** Cunningham, Johnson, et al. (2004).

141 **interacting with a black person . . . cognitive load:** Richeson and Shelton (2003).

141 **automatic settings that tell us how to proceed:** Bargh and Chartrand (1999).

141 **amygdala . . . 1.7 hundredths of a second:** Whalen, Kagan, et al. (2004).

141 **VMPFC . . . decisions involving risk:** Bechara, Damasio, et al. (1994); Bechara, Damasio, et al. (1997); Damasio (1994).

141 **as revealed in their sweaty palms:** Small differences in palm sweat can be detected by passing a small current through the skin, which conducts current more effectively when moist. This technique is known as the measurement of "skin conductance response" (SCR) or "galvanic skin response" (GSR).

142 **Figure 5.2:** Adapted with permission from Whalen, Kagan, et al. and Rathmann (1994).

142 **we need our emotional automatic settings:** Woodward and Allman (2007).

143 **shaped by cultural learning:** Olsson and Phelps (2004, 2007).

# 第三部分 通用货币

## 第6章 绝妙的想法

148 **within-group cooperation:** And also within-group competition.

149 **"moral relativist":** I here refer to *relativism* in the colloquial sense. In philosophy, *relativism* can be rather different. See Harman (1975).

153 **"pragmatism" often has a different meaning:** Pragmatist theories of truth are, roughly, ones according to which claims are true or false depending on the practical effects of believing in them.

154 **make things go as well as possible:** Strictly speaking, I am talking about act consequentialism.

155 **earliest opponents of slavery:** Driver (2009).

156 **two centuries' worth of philosophical debate:** Smart and Williams (1973).

161 **the value that gives other values their value:** Mill (1861/1987), chap. 4, 307–314; Bentham (1781/1996), chap. 1.

161 **"It is better to be a human being dissatisfied":** Mill (1861/1998), 281.

162 **an argument that Mill dashes off in passing:** Ibid., 282–283.

162 **"broaden and build":** Fredrickson (2001).

164 **Measuring happiness:** Easterlin (1974); Diener, Suh, et al. (1999); Diener (2000); Seligman (2002); Kahneman, Diener, et al. (2003); Gilbert (2006); Layard (2006); Stevenson and Wolfers (2008); Easterlin, McVey, et al. (2010).

165 **science of happiness excels:** See previous note.

165 **unemployment is often emotionally devastating:** Clark and Oswald (1994); Winkelmann and Winkelmann (1995); Clark, Georgellis, et al. (2003).

165 **making a bit less money:** Here the debate is between those who say that additional money for the well-off buys no additional happiness (Easterlin, 1974; Easterlin, McVey, et al., 2010) and those who say that it buys some but not much (Stevenson and Wolfers, 2008). At best, happiness seems to increase as a logarithmic function of income, meaning that gaining another unit of happiness requires ten times more income than it took to gain the last unit.

166 **we may soon have such measures:** At least for happiness in the moment. Neural measures of life satisfaction pose a far greater challenge.

166 **This stereotype . . . is undeserved:** Mill (1861/1998), 294.

167 **attempt to outcalculate . . . at our peril:** Ibid., 294–298; Hare (1981); Bazerman and Greene (2010).

169 **down to first principles:** There is a sense in which utilitarians are more closely aligned with ideological collectivists than with ideological individualists. Both utilitarians and ideological collectivists aim for the greater good. The difference is that ideological collectivists are committed to a collectivist way of life as a matter of first principles. In contrast to both utilitarians and ideological collectivists, ideological individualists do not aim for the greater good per se. If some people are foolish and lazy and they get less, that's perfectly fine with individualists, even if their getting less reduces aggregate happiness. For ideological individualists,

the goal is not to maximize happiness but to give people the happiness or unhappiness they *deserve*. The utilitarian take on communism follows the old quip "Great in theory, terrible in practice." Ideological individualists won't even say "Great in theory."

170 **values . . . derive their value from their effects on our experience:** See Sidgwick (1907), 401.

170 **no impact on our experience . . . would not be valuable:** This statement and the one preceding it are not, in fact, equivalent. It could be that all values must have an impact on our experience in order to be valuable, but from this it doesn't follow that the value of a value is derived solely from its impact on our experience. In other words, impact on experience may be necessary for having value, but not sufficient for determining the value of a value.

170 **second utilitarian ingredient is impartiality:** Sidgwick (1907) calls this the axiom of justice.

172 **connection between manual-mode thinking and *utilitarian* thinking:** This association has been challenged by Kahane, Wiech, et al. (2012). They argue that manual-mode thinking favors utilitarian judgments in some cases—like the *footbridge* dilemma—but not in general. To make this point, they conducted a neuroimaging study using a new set of dilemmas in which, according to them, the deontological judgment (the nonutilitarian judgment favoring rights or duties over the greater good) is less intuitive than the utilitarian judgment. However, Joe Paxton, Tommaso Bruni, and I have since conducted an experiment that casts serious doubt on their conclusions, which were not well supported by the neuroimaging data to begin with (Paxton, Bruni, and Greene, under review). We used something called the Cognitive Reflection Test (Frederick, 2005), which can both measure and induce (Pinillos, Smith, et al., 2005) reflective thinking, to test one of their new dilemmas. This is a "white lie" case in which the greater good is served by telling a lie. As a control, we tested one of our standard *footbridge*-like dilemmas. We showed that in *both* cases, being more reflective is associated with more utilitarian judgment. This is a striking victory for the dual-process theory presented here, because it employs a dilemma that both I (Greene, 2007) and these critics thought would work as a counterexample.

173 **the amygdala and the VMPFC:** Actually, the situation is a bit more complicated. Current research suggests that the amygdala functions more like an alarm bell, while the VMPFC is actually more of an integrator of emotional signals, translating motivational information into a common affective currency (Chib, Rangel, et al., 2009). Thus, VMPFC damage may block the influence of automatic settings by blocking the route by which they influence decisions. Decision rules applied by the DLPFC can influence the affective integration in the VMPFC (Hare, Camerer, et al., 2009), but such rules can also be applied without the VMPFC. See Shenhav and Greene (in prep.).

173 **ambiguous Golden Rule:** The Golden Rule is ambiguous because people's situations are always different, and the Golden Rule doesn't tell us which situational differences justify differences in treatment. For almost any disparity in treatment, one can find a formally impartial rule that justifies it: "Yes, and if *you* were king, and *I* were a peasant, then *you* would have the right to do whatever you want to *me*!" The Golden Rule works only when there is agreement about which features of our situations matter morally. In other words, the Golden Rule doesn't set the terms of cooperation. It just says that pure selfishness, as in "I get more just because I'm me," is not allowed. When it comes to resolving conflicts, that's not very helpful, because, as explained in chapter 3, no tribe's values are purely selfish.

## 第7章 寻找通用货币

175 **aware of this problem:** Obama (2006) continued: "Now this is going to be difficult for some who believe in the inerrancy of the Bible, as many evangelicals do. But in a pluralistic democracy, we have no choice. Politics depends on our ability to persuade each other of common aims based on a common reality. It involves the compromise, the art of what's possible. At some fundamental level, religion does not allow for compromise. It's the art of the impossible. If God has spoken, then followers are expected to live up to God's edicts, regardless of the consequences. To base one's life on such uncompromising commitments may be sublime, but to base our policy making on such commitments would be a dangerous thing."

176 **"only people of non-faith can . . . make their case":** See Greenberg (February 27, 2012). Santorum was responding directly to President Kennedy's views, which were similar to Obama's.

176 **Rights . . . *trump* consequences:** Dworkin (1978).

177 **For many modern moral thinkers:** Kant (1785/2002); Hare (1952); Gewirth (1980); Smith (1994); Korsgaard (1996). See also a forthcoming book (still untitled) by Katarzyna de Lazari-Radek and Peter Singer in which they, inspired by Henry Sidgwick, defend utilitarianism as an axiomatizable system.

177 **Other tribes say that earthquakes:** Espresso Education (n.d.).

178 **whether or not it's the moral truth:** What do we mean by "works?" How can we say whether a metamorality "works" without applying some kind of evaluative standard? And how can we apply such a standard without assuming some kind of moral truth or, at least, a metamorality? We'll discuss this problem in more detail later,

but for now the short answer is this: A metamorality "works" if we're generally satisfied with it. And one metamorality works better than another if, in general, we're more satisfied with it. An analogy here is with law. You don't have to believe "Thou shalt not drink under twenty-one" is the moral truth to be satisfied with a law setting the legal drinking age at twenty-one. And different people can be satisfied with such a law for different reasons. General satisfaction with a moral system does not presuppose agreement on moral first principles.

178 **Plato:** Plato's *Euthyphro,* in Allen and Platon (1970).

179 **How can we know God's will?:** Craig and Sinnott-Armstrong (2004).

180 **open letter to Dr. Laura:** The letter is available on many websites in many forms. For one reprinting and a discussion of its origins, see Snopes.com (November 7, 2012).

182 **"Abraham is ordered by God":** Obama (2006).

182 **reflection discourages belief:** Shenhav, Rand, and Greene (2012); Gervais and Norenzayan (2012).

183 **most of them, at least:** Of course, some religions are less tribalistic than others. A decidedly untribal religion is the Unitarian Universalist church.

183 **"pure practical reasoning":** Kant (1785/2002).

183 **views that can't be rationally defended:** Of course, some people *are* committing rational errors, maintaining sets of moral beliefs that are internally inconsistent. But the hard-line rationalist believes that some specific moral conclusions (as opposed to combinations of conclusions) could never be rationally defended. This requires that morality be like math, with substantive conclusions derivable from self-evident first principles. More on this shortly.

184 **manageable set of self-evident moral truths:** I say "manageable" because morality is not like math if the axioms are an enormous set of statements too expansive to be written down.

184 **no one has found . . . axioms:** Why has no one found such axioms? What kind of principles would make good axioms? Given that the axioms need to be self-evident, we might hope for axioms that are "analytic," that is, true by virtue of the meanings of the words used to express them. (The validity of the analytic/synthetic distinction was famously questioned by Quine [1951], but it seems very hard to get by without it [Grice and Strawson, 1956].)

For example, the statement "All bachelors are unmarried" is analytic. You might say that analytic statements are "true by definition," with the caveat that the truth of some analytic statements may be nonobvious, especially if they are long and complicated. It's also possible to have self-evident truths that are not analytic—for example, Euclid's axiom stating that it's possible to connect any two points with a straight line. This is obviously true, but there's nothing in the definitions of "point" and "two" from which one can derive this truth. Put another way, the concept POINT does not *contain* the concept STRAIGHT (or the concept LINE) in the way that the concept BACHELOR *contains* the concept UNMARRIED. Thus, using Euclid's as our model, we might hope to find moral axioms that are self-evidently true but not true by definition. Or we might hope to find axioms that are true by definition. If we're looking for moral axioms, those are our options.

The early-twentieth-century philosopher G. E. Moore (1903/1993) put forth an argument, known as the Open Question Argument, that provides a test for aspiring moral axioms. We can start by thinking of the Open Question Argument as a test for self-evidence, although that's not how Moore thought of it. (Moore thought that propositions that failed the test couldn't be true, but he overlooked the possibility that they could be true but not self-evidently or obviously true.) A test for self-evidence is what we need if we're looking to model morality on math, because, once again, the moral axioms need to be self-evident.

Moore's test works as follows. Take a moral principle that purports to tell you what sorts of things are right, wrong, good, bad, etc. For example, a utilitarian principle like this one:

What's right is what maximizes overall happiness.

If you have utilitarian inclinations, you might think that this principle is not only true but also self-evident. To you, Moore poses the following challenge: Suppose we know that a certain action will maximize overall happiness. Isn't it still an open question whether the action is right? If the answer is yes, then it can't be self-evident that what maximizes happiness is what's right. In this case, you can feel the pull of Moore's Open Question Argument by considering counterexamples. Take the case of pushing the man off the footbridge to save the five. Let's grant that this action will maximize overall happiness. Have we then granted that it's right? Clearly not. The moral question remains open for now, regardless of what we may conclude in the end.

Let's try an even more abstract principle. This one is borrowed, with liberties taken, from Michael Smith (1994):

What's right is what we would want if we were fully informed and fully rational.

This principle may seem to be self-evidently true, but is it really? Suppose we have an action that is favored by someone who is fully informed and fully rational. Is it not an "open question" whether this action is right?

Think, for example, of your classic evil masterminds, such as Hannibal Lecter. Maybe it's true that a baddie like Lecter, who kills and eats innocent people, must be making some kind of logical error, or must be

ignorant of some nonmoral facts. Maybe, but maybe not. The point is that it's not self-evident that this is so. Shaun Nichols has shown that many ordinary people believe that psychopaths know right from wrong but simply don't care, consistent with the idea that one can be morally deficient without being irrational or nonmorally ignorant (Nichols, 2002). Even if we become fully informed and fully rational, it's still an "open question" whether we have therefore become morally perfect. And that means that the principle above can't be self-evident. It may appear to be self-evident (to some of us, at any rate), because we think that being more rational and more informed can only help. But that's very different from saying that full rationality and full information are, self-evidently, all you need for perfect moral judgment and motivation. We could be fully informed and fully rational and still make some moral mistakes. In any case, it's not self-evident that what I just said is false. And that means that the above principle can't be a moral axiom because, even if it's true, it's not self-evidently true.

What's more, even if we were to accept this kind of very abstract principle as an axiom, it wouldn't give us a common currency. It wouldn't tell us how to make trade-offs among competing values. It would simply tell us that the right trade-offs are the trade-offs we would make if we were fully informed and fully rational, which is not much help.

Moore thought that, for any moral principle, the question it purports to answer would always remain open. To see why, we need to think about what a moral principle is. For Moore, a moral principle is one that connects a "natural property" to a "moral property." Take, for example, the principle "Lying is wrong." An action's being an instance of telling a lie is a "natural property" of that action. An action's being wrong is a "moral property" of that action. And what the principle "Lying is wrong" means, in property-speak, is this: If an action has the natural property of being an instance of lying, then it has the moral property of being wrong. What Moore's Open Question Argument suggests is that ascribing natural properties to things will never get us all the way to moral properties, at least not in a way that is self-evident. (Moore's terminology implies that moral properties must be "unnatural," but this is not an essential part of his argument. Instead of speaking about "natural" properties, we can instead speak about apparently factual properties that can be ascribed without controversy, the "facts of the case," as lawyers say.)

Suppose that Joe lied to the police in order to protect his friend. We disagree about whether this is wrong. It is, however, a noncontroversial "fact of the case" that Joe lied. From this fact, are we forced to conclude that what Joe did is wrong? No, says Moore. It's an "open question." Moore argues that all substantive moral principles—not just "Lying is wrong"—have the same limitation. The reason is that all substantive moral principles must span the is-ought gap, with the "natural properties" (the noncontroversial facts of the case) on one side and the "moral properties" on the other. The facts of the case are always about what "is"—facts like the fact that Joe lied. Moral conclusions, in contrast, are always about what "ought" to be, such as the fact that Joe ought not to have lied. Now, it may be true that lying is wrong, but the key point here is that this moral principle can't be self-evidently true. Why? Because even if we agree that Joe lied (a fact about what "is"), it's still an open question whether this act of lying was wrong (a fact about what "ought" to be).

"Lying is wrong" is not self-evidently true, but that's just one candidate. Maybe there are other familiar moral principles that are self-evidently true. We might start with a principle concerning something that seems obviously and unconditionally wrong. How about this: "Torturing kittens is wrong." Isn't it self-evident that it's wrong to torture kittens? Well . . . What if the only way to save a million people is to torture one kitten? Would that be wrong? And is it *self-evident* that it's wrong? Perhaps we need a little tweak: "Torturing kittens is wrong, unless you have a really good reason." But what's a "really good reason"? One that's good enough to justify torturing a kitten? If so, then we have, instead of a substantive moral principle, an empty tautology: "Torturing kittens is wrong unless there is a reason sufficient to justify kitten torturing." We could attempt to be more specific. But how much good does your kitten torturing have to do? And how does the amount of good vary with the amount of torture? And are there answers to these questions that are self-evidently correct? Perhaps, with enough tweaking, we can get a moral statement about kitten torturing (or whatever) that passes the open-question test. But what we're going to end up with is not a moral axiom, a foundational principle from which more specific moral truths can be derived. Rather, it's going to be a very specific and highly qualified statement about the ethics of torturing kittens (or whatever).

We've been talking about the Open Question Argument as a test for self-evidence, and self-evidence is what we need from our axioms. But strictly speaking, a statement can be self-evidently true even if it's being true remains an "open question." How? Consider this statement: All bachelors are not not not not not not not not not not not married. Unless you counted the *not*s, the truth of the previous sentence is, for you, an open question. It turns out that this sentence is true, and it's self-evidently true, meaning that you don't need any evidence beyond what's in the sentence to verify that it's true. Thus, strictly speaking, a statement can be self-evident even if its truth is an open question. And, thus, the Open Question Argument is not, strictly speaking, a test for self-evidence. Rather, it's a test for something more like "obviousness." This leaves open the possibility that there are useful moral axioms that are self-evidently true but not obviously true. What

would such a principle be like? Unlike the "not not" sentence above, such a principle would not be reducible to something simpler. Otherwise the simpler version could be used as the axiom. Thus, such a principle would have to be an irreducibly complex moral statement that can be seen as true without any further evidence or argument, but whose truth is not obvious, due to its complexity. And from this statement (along with others like it, perhaps) and nonmoral facts, we will be able to derive answers to controversial moral questions. I cannot prove that moral axioms of this kind do not exist, but it's fair to say that we should not count on their arriving anytime soon.

In sum, the prospects for modeling morality on math don't look good. To pull this off, we'll need moral axioms that are both self-evident and useful, but there don't seem to be any such axioms around. To be useful, moral principles must enable us to connect the "is" to the "ought," taking us from the "facts of the case" to specific moral answers. And principles that are powerful enough to do that don't seem to be self-evident, however plausible they may be. I cannot prove that morality will never be axiomatized, and therefore made like math. But we best not hold our collective breath.

P.S. Katarzyna de Lazari-Radek and Peter Singer have argued (in an unpublished and untitled book manuscript) that utilitarianism rests on a set of self-evident axioms. Their argument, which is inspired by Henry Sidgwick (1907), is, in my opinion, about as compelling as such arguments get. But in the end I can't agree, for reasons suggested by the foregoing discussion.

186 ***competition between groups*:** Once again, I'm not assuming a group-selectionist account of moral evolution. Here, a "group" may consist of two people, a tribe of thousands, or anything in between.

186 **favor people with genocidal tendencies:** Joyce (2011).

186 **argument, which is controversial, also applies to cultural evolution:** Casebeer (2003). See also Ruse and Wilson (1986) and a rebuttal by Kitcher (1994). Suppose that what evolved biologically was not morality, but a general capacity to acquire cultural practices. And suppose that morality evolved purely culturally, meaning that it is a set of "memes" (cultural variations) that spread because morality outcompeted other memes in the struggle for existence in human brains. In that case, the same argument still applies. The ultimate function of morality would then be to make more copies of itself in the brains of other humans, rather than to make more copies of its associated genes. Moral tendencies might spread simply because they are good at "infecting" brains, rather like catchy tunes and conspiracy theories. Or they might survive because they help their hosts survive. More specifically, moral memes might survive because they help their hosts outcompete the competition. But whatever the case, and whether or not morality does some good along the way, the bottom line of cultural evolution is the spreading of cultural memes, just as the bottom line of biological evolution is the spreading of genes.

186 **the "is-ought" problem:** Hume (1739/1978).

186 **"naturalistic fallacy":** According to G. E. Moore (1903/1993), who coined the term, the naturalistic fallacy is to infer that an entity is good from facts about its "natural" properties—for example, inferring that chocolate is good because it is tasty.

186 **social Darwinists:** The term "social Darwinism" is used primarily as a pejorative. It is often attributed, rather unfairly, to Herbert Spencer. Insofar as Social Darwinism existed as an ideology, it was in the minds of elite capitalists who saw in Darwin confirmation of their preexisting moral and political beliefs. See Wright (1994).

188 **I thought this was *the* question:** Greene (2002).

188 **What really matters . . . path through the morass:** Thanks to Walter Sinnott-Armstrong, Peter Singer, and Simon Keller for pressing me on this point.

188 **Do we call what's left "the moral truth"?:** Here's the dilemma: On the one hand, it seems that some moral views are clearly *better* than others. If a moral position is internally inconsistent, or if its appeal depends on false assumptions, then such a view is in some sense objectively inferior to others that lack such problems. But if we believe in "objectively better" and "objectively worse," then why not believe in "objectively right"? Why not say that the moral truth is just whatever our moral beliefs become after we've objectively improved them as much as possible?

On the other hand, you might think that moral truth requires more than postimprovement agreement. (See Mackie, 1977; Horgan and Timmons, 1992; and Joyce, 2001, 2006.) If what you believe is truly *true*, then it should be impossible for someone to disagree with you without making some kind of objective mistake. Suppose that a well-informed, perfectly rational psychopath—one whose thinking meets our standards for "fully improved"—disagrees with us about, say, the wrongness of torturing kittens. There are two possibilities. First, we might deny that such a person could exist. If his thinking is fully objectively improved, we say, then he must agree that torturing kittens is wrong. But on what grounds can we say that? This is the kind of thing that we can say only if we have direct access to moral truth, which we apparently don't have. Our other option is to accept that a psychopath with fully improved moral thinking (knows all the facts, no internal inconsistencies, etc.) could exist. But this means that someone can reject the moral truth without making

(in any non-question-begging sense) a mistake. But that doesn't sound like *truth*. It's like saying this: "It's true that the sun is larger than the earth, but if you think that the sun is not larger than the earth, you're not necessarily wrong." Huh?

So is there moral truth? In my dissertation (Greene, 2002), I argued that there isn't any moral truth, for the reasons given "on the other hand" above. But now I think that what matters most for practical purposes is the possibility of objective improvement, not the possibility of objective correctness. And this inclines me to say that, for practical purposes, there can be something very much like moral truth (Blackburn, 1993), which might more or less be the moral truth, if not the Moral Truth. But, really, I think it's the wrong question on which to focus, which is why I've paid it relatively little attention in this book. What matters is what we do with the morass, not whether we call the final product the "moral truth."

189 **at the highest level:** "At the highest level" is key. Almost no one thinks that we must be completely impartial in a day-to-day sense, caring for strangers no less than we care for ourselves and our loved ones. But at the same time, we all recognize that, from a moral perspective, we must all be subject to the same rules, even as we occupy different positions within the system established by those rules. If I'm allowed to favor my loved ones over strangers, then so are you, so long as our positions are symmetrical. And if I'm allowed to favor mine more than you're allowed to favor yours, it's because we occupy objectively different positions (e.g., a president of a private club vs. a federal judge). In short, we're allowed to be partial at a low level, but the rules that define when and where partiality is acceptable must be applied impartially.

189 **"moral truth" . . . an open question:** See note "Do we call what's left 'the moral truth'?".

## 第8章 通用货币出现

190 **episode of *The Twilight Zone*:** The episode (Medak, 1986), written by Richard Matheson, was based on his earlier short story and was the basis for the movie *The Box*.

191 ***"If all else is equal"*:** We're assuming that there is no hidden upside here. Breaking your kneecap won't improve your character. You won't meet the love of your life in the hospital. Here, breaking your kneecap is just an unmitigated reduction in your happiness that you can avoid by pushing the button.

193 **go out and check:** At the very least, I expect at least some members of all tribes to follow the utilitarian logic *within the tribe.* There may be some tribes that are so tribalistic that they are essentially psychopathic at the intertribal level. But as noted above, if that's how they are, they are simply not part of the "we" of this conversation. At least not yet. As noted in chapter 3, kindness toward strangers seems to be supported, if not created, by modern market societies (Henrich, Ensminger et al., 2010). In testing this conjecture, methods will have to be tailored to the population(s) being tested.

193 **substantial moral common ground:** You might disagree. You might observe that "if all else is equal," commitments are a dime a dozen. We're all opposed to lying, if all else is equal. We're all in favor of letting people spend their money however they please, if all else is equal. And so on. In other words, many values, perhaps most values, are shared to some extent, and something that we all value to some extent is something to which we're all committed "if all else is equal." Conflict arises primarily because people prioritize different values in different ways. Thus, there's nothing special about our "if all else is equal" commitment to maximizing happiness. We have "if all else is equal" commitments to many, many values. The problem is that all else is never equal.

It is indeed true that "if all else is equal" moral commitments are easy to come by. Nevertheless, our "if all else is equal" commitment to maximizing happiness is special. "No lying" is not a moral system. Nor is "Spend your money however you want," and so on. By contrast, "Maximize happiness" is a moral system. Why is it a system? Because a commitment to maximizing happiness tells us how to prioritize different values—in other words, how to make *trade-offs.* It gives us answers to questions such as "When is it permissible to lie?" and "When does economic freedom go too far?" and so on. Thus, our "if all else is equal" commitment to maximizing happiness is not just a default commitment to one among many moral values. It's a default commitment to what is, or could be made into, a complete system of moral values. That's profoundly important.

194 **objections . . . ultimately driven by *automatic settings*:** The key word here is *ultimately.* I understand that there are some abstract, theoretical arguments against utilitarianism. My claim, however, is that these theories are ultimately motivated by gut feelings. See chapter 11.

196 **make the world that way:** Russell and Norvig (2010).

196 **simple kind of problem solver is a thermostat:** Dennett (1987).

198 **what the human PFC does:** Miller and Cohen (2001).

199 **recognize these errors as errors:** Kahneman (2011).

200 **one way to get . . . *impartiality*:** For developments of this idea, see Gauthier (1987) and Boehm and Boehm (2001) on egalitarianism in hunter-gatherer societies.

200 ***The Expanding Circle:*** Singer (1981).

201 **does not entail abandoning one's *subjective* reasons:** Here, there may be no objective reason to favor one's self, but, for all we've said, there's no objective reason *not* to favor one's self. One might conclude that, objectively speaking, all concerned are equally entitled to be completely selfish.

201 ***empathy*, the ability to feel what others feel:** Batson, Duncan, et al. (1981); Hoffman (2000); Decety and Jackson (2004); de Waal (2010).

201 **to choose one's finger:** Smith (1759/1976), section III.3.4; Pinker (2011), 669–670; and Bloom (in press) make the same point, also citing Smith.

205 **According to John Rawls:** Rawls (1971).

206 **save a life for about $2,500:** Givewell.org (n.d.)

206 **Against Malaria Foundation:** Ibid.

207 **spend that money helping desperately needy people:** Singer (1972, 2009); Unger (1996).

207 **not the only un-splendid implications:** Utilitarianism has other famously counterintuitive implications. First, it fails to distinguish between natural and artificially generated experience (Nozick, 1974). Second, it allows sufficiently large numbers of individuals (e.g., rabbits) with minimally positive experiences to take precedence over many people living good lives, a "repugnant conclusion" (Parfit, 1984). Third, it also allows one individual (a "utility monster") with extremely high-quality experience to take precedence over many people living good lives (Nozick, 1974). I addressed these issues in my undergraduate thesis (Greene, 1997), and Felipe De Brigard (2010) makes a nice, empirically based argument along similar lines concerning the issue of artificial versus real experience. I won't say much about these issues here, because they are, in my estimation, less closely related to real-world moral issues, as they invoke premises that take us deep into the realm of science fiction, pushing the imagination past its emotional, if not conceptual, limits. For further discussion of the utility monster and the repugnant conclusion see note "the 'in principle' version of this objection."

207–08 **an abstract idea . . . specific problems:** For a general discussion of the tension between abstract and concrete thinking, see Sinnott-Armstrong (2008).

208 **hypothetical questions . . . widely underappreciated:** For an excellent discussion of the allergy to hypothetical questions, see Kinsley (2003).

208 **utilitarianism defended:** Two recent popular books—*The Moral Landscape,* by Sam Harris (2010), and *The Righteous Mind,* by Jonathan Haidt (2012)—discuss recent advances in moral psychology/neuroscience and end up favoring a version of utilitarianism, as I do in this book. Haidt does not attempt to defend utilitarianism against the objections listed above. Harris explains why utilitarianism's foundational principles are reasonable, as I do here in part 3, and as Bentham and Mill did long ago. Harris, however, pays little attention to the many compelling objections to utilitarianism listed above. Harris aims to show that science can "determine human values," but I don't think he does this, at least not in the sense that has been controversial among moral philosophers. He shows that, given an assumption of utilitarian values (which is neither supported nor undermined by science), science can determine further values. In other words, science can help us figure out what makes people happy. I'm sympathetic to Harris's practical conclusions, but in my opinion he has ignored, rather than solved, the problem he seems to want to address. Many have offered similar assessments of his book, and he has responded. See Harris (January 29, 2011). In part 4, I attempt to give utilitarianism a more thorough (though inevitably incomplete) defense, drawing on the new science of moral cognition.

# 第四部分 道德信念

## 第9章 值得警惕的事件

211 ***accommodation* and *reform*:** My use of these terms comes from Brink (2011). See also Bazerman and Greene (2010) on utilitarian accommodation.

211 **greater good . . . in the long run:** This still leaves us with the problem of endorsing such actions in principle, a problem that I take seriously. More on this later.

212 **apparent intellectual inferiority of women:** Mill (1895).

212 **too inflexible to serve as the ultimate arbiters:** These arguments resemble and build on ones made previously by Jonathan Baron (1994), Cass Sunstein (2005), Peter Singer (2005), Walter Sinnott-Armstrong (2004), and Stephen Stich (2006), among others. See also Greene (1997, 2002, 2007, under review).

212 **race of the defendant:** Baldus, Woodworth, et al. (1998); Eberhardt, Davies, et al. (2006). See also US General Accounting Office (1990).

214 **controlled for people's real-world expectations:** In these studies (Greene, Cushman, et al., 2009) we asked people versions of these three questions: In the real world, what are the odds that this attempt to save

five lives would go as planned? What are the odds that it would go worse than planned? What are the odds that it would go better? We then used people's answers to these questions to statistically control for people's real-world expectations. In other words, we asked whether we can predict people's judgments simply by knowing their real-world expectations. What we found was that we could a little bit, but not very much. It seems that when people say no to trading one life for five in these cases, it's not because of their real-world expectations. It's primarily because of the features of the dilemmas described below.

215 **"personalness" of the harmful action:** Greene, Cushman, et al. (2009). The meaning of "personal" that comes out of these more recent studies is different from the one tentatively proposed earlier (Greene, Sommerville, et al., 2001).

216 **what seems to matter is touching:** See also Cushman, Young, et al. (2006); Moore, Clark, et al. (2008); and Royzman and Baron (2002).

216 **without touching:** Of course, there's a sense in which this involves touching, namely touching with a pole.

216 **a big drop:** Greene, Cushman, et al. (2009).

217 ***accommodation*:** That is, utilitarianism can accommodate the fact—I assume it's a fact—that a willingness to engage in personally harmful utilitarian actions likely indicates a more general antisocial willingness to harm people. See Bartels and Pizarro (2011) on Machiavellian utilitarianism. Conway and Gawronski (2012), however, show that the Machiavellians are not really utilitarian but rather undeontological.

218 **forbidden by international law:** McMahan (2009).

218 **American Medical Association:** American Medical Association (1991)

220 **no to the *footbridge* case:** Greene, Cushman, et al. (2009).

220 ***loop* case:** Thomson (1985).

221 **no point in hitting the switch:** If you're thinking that this buys the five workmen more time, we can bulge out the main track in the other direction.

221 **81 percent . . . approved:** This result is consistent with Thomson's (1985) intuition, and with Waldmann and Dieterich (2007), but not with Hauser, Cushman, et al. (2007). See Greene, Cushman, et al. (2009) for an explanation.

221 **"Doctrine of Triple Effect":** Kamm (2000).

222 **87 percent . . . approved:** As of this writing, these data have not been published. This experiment was conducted concurrently with those reported in Greene, Cushman, et al. (2009), using identical methods. Testing materials and data available by request.

222 **magic combination:** If you've been paying really close attention, you'll have noticed a gap in this pattern. The *collision alarm* case gets 86 percent approval, and the *remote footbridge* case gets 63 percent approval. And yet, these are both means cases without personal force. Why the difference? It seems that there is another factor that interacts with the means/side-effect distinction: whether or not the victim gets dropped from a footbridge. If you do a dropping version of the *collision alarm* case, the approval ratings drop to roughly the level of *remote footbridge*. But if you do a dropping version of the *switch* case, the drop has little to no effect. More generally, it seems that multiple "forcey" factors (force of muscles, force of gravity) interact with the means/side-effect factor. Even more generally, it seems that the effect of the means/side-effect factor does not depend entirely on the presence of personal force. But it does depend on the presence of some other factors, as demonstrated by the *collision alarm* and *loop* cases.

223 **bound up with . . . personal force:** But not completely bound up. See previous note.

223 **no knowledge of the doctrine:** Cushman, Young, et al. (2006); Hauser, Cushman, et al. (2007).

223 **intuitions that justify the principle:** Cushman and Greene (2011).

224 **when we contemplate harming:** This idea is similar to Blair's earlier idea of a violence inhibition mechanism (Blair, 1995). Cushman (in press) has a model according to which the emotional response to intentional personal harm is triggered not by a dedicated alarm system but by a learned negative emotional response encoded within a more general emotional learning system (more specifically, a "model free" learning system). Cushman's model preserves the critical features of what I'm here calling the myopic module. First, the emotional response is blind to side effects (myopia) and, for the reasons given here, related to the analysis of action plans. Second, this system's internal operations are not accessible to introspection. That is, they are "informationally encapsulated" (modularity). But if Cushman is right—and I suspect that he is—this system is not specifically dedicated to serving the function that it serves here. That said, one can think of this learned association as a kind of acquired module.

225 **as Hobbes observed:** Hobbes (1651/1994).

225 **smash his head in with a rock:** It's not clear whether other species, such as chimps, can engage in this kind of premeditated violence. Chimps go on raiding parties, killing members of neighboring groups, but whether these killings are performed with a conscious goal in mind is unclear. They may be more like animal migrations—functional, complicated, and socially coordinated but not consciously carried out with a purpose.

226 **dangerous for . . . the attacker:** DeScioli and Kurzban (2009).

226 **how you treat others:** Dreber, Rand, et al. (2008).

226 **contemplating an act of violence:** Blair (1995).

227 **at least somewhat "modular":** Fodor (1983).

228 **Mikhail . . . Goldman . . . Bratman:** Mikhail (2000, 2011); Goldman (1970); and Bratman (1987).

230 **approving more of . . . harmful side effects:** Schaich Borg, Hynes, et al. (2006); Cushman, Young, et al. (2006).

233 **sounds the alarm:** That is, the *between-dilemma* variability is determined by the strength of the automatic responses. But in fact, the evidence suggests that much of the *within-dilemma* variability is determined by individual differences in reliance on manual mode. See Paxton, Ungar, and Greene (2011) and Bartels (2008).

234 **adding some pushing to the loop case:** At the same time, adding a push to *loop* does appear to have *some* effect, which complicates things for the modular myopia hypothesis. Adding a push to *switch* has little or no effect, and, ideally, the same would be true for adding a push to *loop*. We do see an effectfully consistent with the modular myopia hypothesis when it comes to adding *drops* (dropping onto the tracks from a footbridge through a trapdoor). That is, adding a drop to *switch* has little or no effect, and adding a drop to *loop* has little or no effect, but adding a drop to *collision alarm* significantly lowers approval ratings. This is all a work in progress, and it is not entirely clear to me what is going on. The critical point, for now, is that personal force and dropping do not seem to supply a fully adequate account of the gap between *footbridge* and *loop*. For now, I'm putting these unresolved ambiguities aside, because my purpose in this section is to lay out the modular myopia hypothesis as a *hypothesis* and not as a theory that currently explains everything.

236 **why . . . chain with the harm . . . *secondary* one?:** The secondary causal chain is secondary because it is parasitic on the primary causal chain. The turning of the trolley away from the five makes sense as a goal-directed action all by itself, without reference to the secondary causal chain, that is, to what happens after the trolley is turned. But the secondary causal chain cannot stand alone. This is because the secondary causal chain, to make sense as a complete action, must extend all the way back to the body movement, which is the hitting of the switch. But the hitting of the switch makes sense only with reference to the primary causal chain, that is, to the fact that the unturned trolley will proceed down the main track and kill the five if nothing is done.

236 **a means that is *structured like a side effect*:** Kamm (2000) refers to this kind of structure as a case of "triple effect," one in which there is a foreseen event that is recognized as causally necessary for the achievement of the goal and yet is, in some morally relevant sense, not intended.

238 **cost-benefit . . . sufficiently compelling:** Nichols and Mallon (2006); Paxton, Ungar, and Greene (2011).

239 **nested multitasking:** Koechlin, Ody, et al. (2003).

240 **"Doctrine of Doing and Allowing":** Howard-Snyder (May 14, 2002).

242 **choose between pairs of objects:** Feiman et al. (in prep.).

242 **infants' brains represented . . . experimenter wanted:** This effect was first demonstrated by Woodward and Somerville (2000).

244 **people evaluated both active and passive harmful actions:** Cushman, Murray, et al. (2011).

244 ***ignoring* . . . requires more manual-mode DLPFC activity:** An earlier study provided more ambiguous evidence. Cushman, Young, et al. (2006) had people evaluate harmful actions and harmful omissions and then justify their ratings. About 80 percent of the time, people who distinguished between actions and omissions in their ratings were able to justify their judgments with an explicit appeal to the action/omission distinction. However, this means that about 20 percent of these people did what they did without knowing what they were doing. Clearly, these people were not consciously applying the action/omission principle in manual mode. It's not clear whether some or all of the 80 percent were doing so, or if they became conscious of the action/omission principle only after drawing the distinction intuitively. The brain-imaging data suggest the latter.

244 **tongue, fingers, and feet:** Hauk, Johnsrude, et al. (2004).

245 **jacking up the price . . . feels less bad if done indirectly:** Paheria, Kasam, et al. (2009).

246 ***specifically intended*:** There is a technical problem with calling any of these harms "specifically intended." For example, in the *footbridge* case, one might say that what is specifically intended is using the man's body to block the trolley, which does not *logically* entail any harm to the man. (What if it's Superman?) By this interpretation, the death of the man and the pain he experiences are just contingent *side effects*, unfortunate by-products of using the man's body to stop the trolley. While this interpretation is possible in principle, this is clearly not how our brains represent these events. Thus, there is an interesting psychological problem here: namely, to understand the mechanism that parses events in these contexts.

247 **explanation of our sensitivity to the action/omission distinction:** Cushman, Young, et al. (2006) have found effects of means versus side effect for passive harms, but these are likely cases in which the omission is unusually purposeful, very specifically failing to do what one would ordinarily do in order to save more lives. Thus, it may be possible, but unusual, to have omissions as part of an action plan, as in a recipe ("Do not remove the fritters from the oven until they are golden brown"). Also, it's worth noting that the means/side-effect effect is much weaker for the omission cases.

247 **represent causes in terms of *forces*:** Talmy (1988); Wolff (2007); Pinker (2007).

247 **hitting, slapping, punching:** This doesn't mean that it can't learn to respond to other kinds of violence, such as gun violence. It's possible that guns are sufficiently familiar that we incorporate the explosive force of a gun into the body schema, conceptualizing it as a force that we personally control. The same may also happen with gravity. These are interesting empirical questions. For a fascinating and, I predict, very important theory of how we learn to recoil at certain kinds of harms, I recommend Cushman ("Action, Outcome").

247–48 **hard . . . to think of actions that don't feel violent:** The best candidate that I can think of is surgery, but surgery *does* feel violent. It's just that surgeons learn to get over that feeling (if they're not psychopathic), and we don't blame them for what they do, because we know that their actions are for the good of the patient.

248 **millions of people can be saved by pushing:** Paxton, Ungar, and Greene (2011). See also Nichols and Mallon (2006). The 70 percent figure comes from unpublished data using the same methods as Greene, Cushman, et al. (2009).

248 **would all be more psychopathic:** Bartels and Pizarro (2011); Glenn, Raine, et al. (2009); Koenigs, Kruepke, et al. (2012).

249 **inferences about moral character:** Pizarro, D.A. and Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & Shaver, P. (Eds) The Social psychology of morality: Exploring the causes of good and evil. APA Press.

253 **abortion and the Doctrine of Double Effect:** Foot (1967).

253 **If harming the environment felt like pushing:** See also Gilbert (July 2, 2006).

# 第10章 公正与公平

255 **a dollar spent in the right way:** Givewell.org (n.d.); Sachs (2006); Singer (2009).

255 **more accountable than ever:** www.givewell.org.

256 **"decency to admit that I'm a hypocrite":** Exchange described by Simon Keller.

258 **neighbors are already doing it:** Cialdini (2003).

258 **moral problem as . . . originally posed:** Singer (1972). I've made some minor modifications to Singer's thought experiment.

259 **hard to justify treating the . . . faraway starving child differently:** Jamieson (1999).

260 **"trolleyology" of Peter Singer's problem:** Manuscript in preparation, based on Musen (2010). Much of the work done in these experiments was based on thought experiments carried out by Peter Singer (1972) and Peter Unger (1996).

261 **heavily influenced by mere physical distance:** Nagel and Waldmann (2012) claim that mere physical distance does not matter and that the relevant factor is informational directness. However, my experiments with Musen show effects of mere distance independent of informational directness. In any case, it would be hard to argue that informational directness per se is a normatively significant factor. Therefore, the main conclusions reached here would not change if Nagel and Waldmann's conclusions are correct.

261 **faraway starving children don't:** Note that in Trolleyland, spatial distance didn't seem to matter much, whereas here it does. This is likely because we're dealing with a different automatic setting, one that responds to preventable harm rather than actions that cause harm. It could also be because the distances in these two types of cases differ by at least two orders of magnitude.

262 **from cultural evolution:** Pinker (2011); Henrich, Ensminger, et al. (2010).

262 **unidentified "statistical" people:** Some aid organizations deliberately pair individual donors with individual recipients in order to make the experience more personal.

262 **economist Thomas Schelling:** Schelling (1968).

262 **"Baby Jessica":** Small and Loewenstein (2003); *Variety* (1989).

262 **"statistical death":** Schelling (1968); Small and Loewenstein (2003).

263 **identifiable . . . "statistical," victims:** Ibid.

264 **one sick child . . . or eight:** Kogut and Ritov (2005).

264 **numbers as small as *two*:** Slovic (2007). Note: My recommendations mirror Slovic's specific suggestions about how we might change our approach to the world's needy.

265 **give us legitimate moral obligations and options:** Smart and Williams (1973).

265 **lives defined by relationships . . . must take this . . . into account:** See Sidgwick (1907), 434.

266 **counterproductive to pooh-pooh philanthropists:** Note: See Sidgwick (1907), 221, 428, and 493.

268 **if only they knew:** A. Marsh, personal communication, January 31, 2013.

268 **Wesley Autrey:** Buckley (January 3, 2007).

268 **only human:** Note: See Parfit (1984) on blameless wrongdoing, 32.

268 **utilitarian rationale for punishing people:** Bentham (1830).

268 **favored by many:** Kant (1785/2002).

268 **a little extra justice before pushing off:** Cited in Falk (1990), 137.

270 **Prisoners are frequently sexually abused:** Mariner (2001); Gaes and Goldberg (2004).

271 **is barbaric, we say:** You might object that this is not a fair comparison, because prison rape as it occurs now is a chancy affair, while state-sanctioned rape would be a sure thing. Fair enough. We can use a roulette wheel to introduce an element of chance into our official state raping policy. Now do you like it?

271 **our current criminal justice system, which is highly retributive:** Tonry (2004).

272 **natural function of punishment is quasi-utilitarian:** I say "quasi-utilitarian," however, because our sense of justice is not necessarily designed to make us happier. The Us who benefits from punishment does not necessarily include everyone, and our sense of justice does not necessarily weight everyone's well-being equally. Still, overall, the existence of punishment is a good thing, by utilitarian standards or any reasonable standard.

272 **obvious utilitarian answer:** Carlsmith, Darley, et al. (2002).

272 **punishing based solely on how they *feel*:** Baron and Ritov (1993); Carlsmith, Darley, et al. (2002); Kahneman, Schkade, et al. (1998).

273 **Crimes with lower detection rates:** Carlsmith, Darley, et al. (2002).

273 **people punished "determined" transgressors about twice as much:** Small and Loewenstein (2005).

273 **presented . . . with . . . "deterministic" universe:** Nichols and Knobe (2007).

274 **abstract judgment goes out the window:** See also Sinnott-Armstrong (2008).

275 **maximizing happiness can lead to gross injustice:** Rawls (1971), 158–161.

276 **one-to-one ratio is a *conservative* assumption:** The more typical situation, historically, is that one slaveholder owns many slaves. This only makes it harder for slavery to maximize happiness. One could imagine, however, going in the opposite direction, with a "time share" arrangement in which, say, five people own a slave collectively. But this doesn't change the fundamental math described below. This would mean, let's say, that each slave-share owner gets the equivalent of $10,000 additional income per year. Would you be willing to spend a fifth of your life as a slave in order to get a $10,000 raise?

277 **additional income . . . adds relatively little to one's happiness:** Easterlin (1974); Layard (2006); Stevenson and Wolfers (2008); Easterlin, McVey, et al. (2010). As noted earlier (see chapter 6, "making a bit less money"), the debate is about whether additional income for those already well-off adds nothing, or relatively little, to one's happiness.

277 **not a tentative finding:** Rawls (1971, 158–161) suggests that it is. But he was writing before research on happiness took off.

278 **"utility monsters":** Nozick (1974).

278 **no goods to be extracted . . . outweigh the horrors:** Still not convinced? Let's try a bit harder to think of realistic examples of utility-maximizing oppression. What about the old transplant case (Thomson, 1985)? What if a single healthy body could provide lifesaving organs to twenty people? Would a utilitarian allow us to randomly kidnap people and kill them for their organs, assuming this maximizes happiness? No, because there are clearly better alternatives. Before resorting to kidnapping, which would cause widespread panic and grief, we could establish a legal market for organs. You may or may not think this is a good idea, but it's not grossly unjust, à la slavery. Reasonable people can disagree about whether there ought to be a well-regulated market for human organs.

What about the kind of oppression in which oppressing one person can benefit thousands of people? What about the crowds who cheer as lions tear into the entrails of hapless gladiators? Or people who enjoy child pornography? If there are enough gleeful onlookers, can such suffering be justified? Only if you think it's a net gain when people take joy in the exploitation and suffering of others. We can imagine a hypothetical world in which taking joy from the suffering of innocent people has no detrimental effects, but that's not the real world.

280 **easy to mistake utilitarianism for "wealthitarianism":** Greene and Baron (2001).

281 **counting dollars, rather than happiness:** Rawls (1971, 158–161, 167–168) is aware of the argument that people exhibit diminishing marginal returns of utility from goods and that therefore utilitarian policies will tend to favor egalitarian outcomes. But he dismisses this argument as providing insufficient moral assurance. In doing this, he makes two assumptions. First, even if utilitarianism is generally egalitarian, it will sometimes favor social inequality. Second, sometimes the inequalities that utilitarianism favors will be morally repugnant, à la slavery. Rawls is right about the first assumption (see above), but he takes this first assumption as license to make the second assumption, which is highly doubtful, at least in the real world. Rawls thinks the second assumption is also reasonable, but that, I claim, is because he's making the same error as everyone else: confusing utility with wealth.

283 **pattern . . . predicted . . . based on reading Rawls:** Rawls's error is committed most starkly when he argues (Rawls, 1999, 144) that people *should* be risk averse with respect to utility, implying that some utility is worth more (i.e., has more utility) than other utilities.

284 **the "in principle" version of this objection:** Suppose that it really were possible to maximize happiness by oppressing some people. Wouldn't that still be wrong? And doesn't that show that there's something rotten at the core of utilitarianism? Here the classic example is Nozick's (1974) utility monster, which I alluded to above. Nozick's utility monster gains enormous quantities of happiness by eating people. But it seems that it would be wrong to feed innocent people to the utility monster, even if doing so would, by hypothesis, maximize happiness. Another famous case comes from Derek Parfit (1984), who imagines a choice between two types of worlds: one in which very many people are very happy and one in which many, many more people lead lives that are "barely worth living." The "repugnant conclusion" that follows from utilitarianism is this: No matter how good the world is, there is always a better world, consisting of many, many more individuals whose lives are only minimally good. To drive the point home further, one can even substitute animals for people, so long as we agree that animal experience counts for something. One can imagine an immense warehouse full of trillions of rabbits whose brains are hooked up to stimulators that intermittently produce mild levels of rabbit gratification. What each rabbit gets is not that great, but there are *so many rabbits*. Thus, utilitarian revolutionaries could, in principle, justify destroying our world in order to realize their dream of building an enormous rabbit gratification factory. This prospect strikes most people as unjust. (Of course, no one has bothered to ask the rabbits!)

I have two responses to these "in principle" objections. First, once again, I'm not claiming that utilitarianism is the moral truth. Nor do I claim that it perfectly captures and balances all of human values. My claim is simply that it provides a good common currency for resolving real-world moral disagreements. If the utility monsters and the rabbits ever arrive, demanding their utilitarian due, we may have to amend our principles. Or maybe they would have a good point, albeit one that we have a hard time appreciating.

Which brings me to my second response to these "in principle" objections. We should be very wary of trusting our intuitions about things that defy intuitive comprehension. The utility monster and the rabbits both push our intuitive thinking beyond its limits. More specifically, they push along orthogonal dimensions: *quality* and *quantity*. The utility monster is a single individual (low quantity) with an incomprehensibly high *quality of life*. He gets more out of a single meal than you get out of your entire existence. The rabbits, by contrast, have a rather low quality of life, but the *quantity* of rabbits defies intuitive comprehension. Of course, there is a sense in which we can understand these things. After all, I just described them to you, and you understood my description. But it's your *manual mode* that is doing the understanding. You cannot understand *intuitively* what it's like to eat a meal that produces more happiness than an entire happy human lifetime. Likewise, you can't *intuitively* distinguish between a million rabbits and a trillion rabbits. We can think about these things in an abstract way, but asking us to have gut feelings about such things is like asking a bird to imagine a worm that's a mile long.

284 **Rawls's argument from the "original position":** If you know your Rawls, you'll have noticed that I didn't actually address his official argument against utilitarianism. Rawls argues that the most just organizing principles for a society are those that people would choose from behind a "veil of ignorance," not knowing which positions in society they will occupy. And he argues that people in this "original position" would not choose a utilitarian society, because the possible downside of living in a utilitarian society is too great. In other words, Rawls's official argument depends on the same mistaken assumption described above, which is that a utilitarian society could be oppressive in the real world, with human nature as it actually exists. Rawls's argument also involves some serious fudging related to risk aversion and the structure of the original position. For more on this, see chapter 11, "central argument in A Theory of Justice. . . ."

# 第五部分 道德出路

## 第11章 深度实用主义

289 **Ten percent . . . control 70 percent:** Davies, Shorrocks, et al. (2007). See also Norton and Ariely (2011).

289 **Alex Kozinski:** Alex Kozinski and Sean Gallagher, "For an Honest Death Penalty." *New York Times*, March 8, 1995.

292 **second moral compass:** Thanks to Scott Moyers for suggesting the "two compasses" metaphor.

295 **how the brain gets out of this pickle:** Botvinick, Braver, et al. (2001). This theory is somewhat controversial, but that need not concern us here. Our interest is in the cognitive strategy, regardless of whether the brain actually uses it. That said, I know of no alternative solution to the regress problem described above.

296 **engage the ACC and DLPFC:** Greene, Nystrom, et al. (2004); Greene and Paxton (2009); Cushman, Murray, et al. (2011).

297 **wiser when we acknowledge our ignorance:** Plato (1987).

297 **"illusion of explanatory depth":** Rozenblit and Keil (2002); Keil (2003).

297 **applied this idea to politics:** Fernbach, Rogers, et al. (in press).

297 **left their strong opinions intact:** The demand for reasons did moderate some people's views, but these tended to be people who couldn't produce any reasons at all when asked.

297 **may even do the opposite:** Tesser, Martin, et al. (1995).

297 **an alternative approach to public debate:** Sloman and Fernbach (2012); Fernbach (May 27, 2012).

298 **men crossing two different bridges:** Dutton and Aron (1974).

298 **make up a plausible-sounding story and go with it:** For a classic demonstration of this kind of interpretive effect, see Schachter and Singer (1962).

298 **not an isolated phenomenon:** Bargh and Williams (2006); Wilson (2002).

298 **choose . . . panty hose:** Nisbett and Wilson (1977).

299 **"change into my work clothes":** Stuss, Alexander, et al. (1978).

299 **"split-brain" patients:** Gazzaniga and Le Doux (1978).

300 **plausible narrative:** Bem (1967); Wilson (2002).

300 **consummate moral rationalizers:** Haidt (2001, 2012).

300 **"Concerning Wanton Self-Abuse":** Kant's "Concerning Wanton Self-Abuse" is a section in the *Metaphysics of Morals* originally published in 1797. See Kant (1994).

301 **using someone as a means:** Ibid.

301 **"*Kant's joke*":** Nietzsche (1882/1974).

301 **"born slaves":** Bernasconi (2002).

301 **Rationalization . . . enemy of moral progress, and thus of deep pragmatism:** The argument made in this section, along with other parts of this chapter, was originally made in Greene (2007).

302 **"outweighed" by the rights of the five:** Thomson (1985, 1990). Note that Thomson has changed her mind and now thinks that it's wrong to turn the trolley (Thomson, 2008). What this essentially does is put the rights theorist's explanatory burden on the act/omission distinction—that is, unless one thinks that we're obliged to actively turn the trolley away from the one and onto the five.

303 **we have no duty to save them:** Jamieson (1999).

303 **The rights and the duties follow the emotions:** Kahane and colleagues (2010, 2012) have argued that there is no special relationship between automatic emotional responses and characteristically deontological moral judgments, and that the appearance of such a relationship is the product of a biased selection of stimuli. For evidence to the contrary, see chapter 6, "connection between manual-mode thinking," on the "white lie" case, and Paxton, Bruni, and Greene (under revision).

303 **sexiness is in the mind of the beholder:** This doesn't mean that perceptions of sexiness are *arbitrary*. As evolutionary psychologists have pointed out (Miller and Todd, 1998), what we find sexually attractive is typically indicative of high reproductive potential. But the fact that sexual attraction is non-arbitrary and biologically functional does not imply that it's *objectively correct*. There's no meaningful sense in which we're objectively (absolutely, nonrelatively) correct about who's sexy while baboons are objectively incorrect—or vice versa.

304 **cognitive apparatus . . . concrete objects and events:** Lakoff and Johnson (1980).

304 **nonnegotiable facts:** Of course, there are facts about which rights and duties are granted by law, but in the midst of a moral controversy, such legal facts almost never settle the question. Public moral controversies are about what the law *ought to be*, not about what it is.

305 **not arguments, but weapons:** In a provocative paper, Mercier and Sperber (2011) claim that reasoning is just one big weapon for persuading others to do what we want. This strikes me as highly implausible. What makes their argument go is that they exclude from the category of "reasoning" all of the boring, everyday things for which we use our manual modes, such as figuring out the best order in which to run one's errands. ("I'd better go food shopping last or else the ice cream will melt in the car.") This argumentative theory of reasoning also makes little evolutionary sense. Reasoning did not emerge de novo in humans. Indeed, the neural structures that we use for reasoning are the same ones that our primate relatives use to solve their own (fairly) complex problems. However, chimps and macaques very clearly do not engage in persuasive verbal jousting.

306 **Dershowitz once told a handful:** This was told to me and other undergraduates at a "meet the professor" lunch in 1994. The details are as close as I can recall.

306 **costs of lavishing time and attention:** To spell out this point more explicitly: Dershowitz's response was clever because it distinguished the benefits from the costs. He essentially said: I'm not refusing to debate you because I'm afraid. I'm refusing to debate you because there are costs associated with taking cranks like you seriously. But if you're willing to debate me in a way that denies you the credibility that you seek (the cost), then I'm happy to have a free exchange of ideas (the benefit).

307 **good . . . to reject some ideas out of hand:** See also Dennett (1995) on "good nonsense."

309 **truly moral considerations on both sides:** Here, by "truly moral" I mean not just tribalistically moral.

309 **those of Peter Singer:** Singer (1979), chap. 6; Singer (1994).

310 **manual-mode scrutiny:** For arguments along the lines of those made here, see Singer (1979), chap. 6; Singer (1994).

310 **late-term abortions are not:** If you're okay with late-term abortion, what about infanticide? Most of the arguments below apply equally well.

310 **Both early- and late-term abortions prevent a human life:** The odds of living may differ, but surely this difference in the odds of successful birth (say, 60 percent vs. 95 percent) can't be the difference between having a right to life or not. And what if a late-term fetus, for some reason, had the same odds as a typical early-term fetus? Now would it be okay to abort it?

310 **Viability is as much a function of technology:** You might say that what matters is the ability to survive without special technology. If that's right, then what about a nine-month-old fetus that, due to an atypical medical condition, can survive outside the womb, but only with the temporary help of readily available technology? Is it okay to abort that late-term fetus simply because it's not viable without technology?

310 **born as early as twenty-two weeks can survive:** Stoll, Hansen, et al. (2010).

310 **thanks to new technology . . . first-trimester abortions have become immoral?:** Perhaps you're inclined to say yes. After all, you might think that the possibility of being kept alive from that stage of development is morally significant. But note that it's not that technology enables the fetus to survive from that stage of development. Rather, technology enables the fetus to survive from that stage *outside the womb.* The fetus is already able to survive from that stage of development without the fancy technology. It just has to stay inside the womb! We already have the "technology" to keep early-term fetuses alive.

311 **more than being a moral vegetarian:** At least it requires giving up certain kinds of meat. One could perhaps make room for certain other kinds of meat, such as that coming from animals that lack the relevant features.

311 **pro-choicers are unwilling to go that far:** And even then, it's not clear that the argument works. Animal rights activists typically focus on the suffering that animals experience while being raised for food. If that's the reason why it's wrong to eat meat, then the same argument would not apply to late-term abortion, as long as the abortion process did not involve suffering, or did not involve a lot of suffering.

312 **Deanna Troi . . . not human:** Okay, okay: not *fully* human. She's only half Betazoid. But the point applies to her maternal relatives. *Jeez*.

312 **can move their bodies:** Dongen and Goudie (1980).

314 **which human shall be, if anyone is to be:** Of course, with only one sperm on deck, the odds of fertilization are lower, but so what? Pro-lifers would not allow us to abort a zygote simply because it has, for whatever reason, low odds of surviving.

314 **robbed an innocent person of his life?:** Moreover, the idea that conception determines identity seems to have more to do with our limited knowledge than with facts about what has or has not been determined. When a couple sets out to conceive a child the old-fashioned way, they may not know which sperm and egg are going to join, and we may have no way of knowing. But whichever child is going to result from a given act of sexual intercourse, *that's* the child that's going to result. And if the couple decides not to go through with it, it's *that child* who will not exist as a result of their backing out. When this happens, no one knows, or will ever know, who "that child" is, but so what? If their having sex would have led to some specific child's existing, then their refraining from having sex led to the nonexistence of that specific child. (I'll refrain from getting into the problem of determinism at this point.)

314 **full biology lesson:** Gilbert (2010), 6, 14, 123–158, 301.

316 **"I think even if life begins in that horrible situation of rape":** Madison (2012).

317 **campaign went up in flames:** Haberkorn (2012).

317 **"problem of evil":** Tooley (2008).

318 **silent drama:** Heider and Simmel (1944).

318 **attributions happen so automatically:** Heberlein and Adolphs (2004).

319 **animals . . . move and have eyes:** Many of us would have a hard time killing the animals we eat, but that's probably just because we're not used to it. Our ancestors did this for millions of years.

320 **Most tribes believe in souls:** Bloom (2004).

327 **what's right and wrong when it comes to life and death:** Beauchamp, Walters, et al. (1989); Baron (2006); Kuhse and Singer (2006).

327 **less optimistic . . . about . . . sophisticated moral theory:** See also Greene (under review).

328 **"reflective equilibrium":** Daniels (2008).

328 **"considered judgments":** Gut reactions are not the same as "considered judgments," but they play a dominant role in determining them.

330 **Aristotle . . . great champion of common sense:** Aristotle (1941).

330 **Aristotle is essentially a tribal philosopher:** MacIntyre (1981).

330 **Aristotelian virtue theory . . . revival:** As part of the Aristotelian revival, I include not just virtue ethics proper (Crisp and Slote, 1997; Hursthouse, 2000), but also "sensibility" theories (Wiggins, 1987),

particularism (Dancy, 2001), and the like—all of the approaches to normative ethics that have given up on discovering or constructing explicit moral principles that tell us what to do. The revival is due in large part to Alasdair MacInytre (1981), who has a similar diagnosis of modern moral problems but thinks that a revamped form of virtue theory is the best we can do, following on the failures of Enlightenment moral theory.

332 **cannot be "universalized":** Kant (1785/2002).

332 **Kant's argument requires impossibility:** Kant's universalization argument is not simply a version of the familiar "What if everyone did that?" argument. He's not merely saying that it would be *bad* if everyone were to lie, break promises, et cetera. That's a *utilitarian* argument against lying—rule-utilitarian or act-utilitarian, depending on how you interpret it. This isn't good enough for Kant, because he wants an absolute prohibition against lying, one that doesn't depend on how things happen to work out in the real world. (Things tend to go badly if everyone lies, et cetera.) He wants morality to be like math: necessarily true and knowable with certainty. See Korsgaard (1996), chapter 3.

333 **well aware of the flaws . . . they have their replies:** See, for example, Korsgaard (1996).

333 **central argument in *A Theory of Justice* . . . essentially a rationalization:** While there is much to admire about Rawls's work, and the man himself, I believe that his central argument is essentially a rationalization, an attempt to derive from first principles the kind of practical moral conclusions that he intuitively favors, which he mistakenly believes to be at odds with utilitarianism. (See pp. 279–84.) Rawls's main argument is laid out in chapters 1, 2, and 3 of *A Theory of Justice* (1971).

I mentioned earlier that utilitarianism begins with two very general moral ideas. First, happiness is what ultimately matters and is worth maximizing. Second, morality must be impartial. Essentially, Rawls keeps the impartiality assumption but drops the assumption that happiness is what ultimately matters. He replaces the assumption that happiness is inherently valuable with the assumption that *choice* is inherently valuable. Thus, for Rawls, the best organizing principles for a society are the ones that people would choose if they were to choose impartially. This is a great idea, with roots in the philosophies of both Kant and John Locke. (Rawls, like Locke, is a "contractarian.")

So how do we figure out what people would choose if they were to choose impartially? To answer this question, Rawls constructs a thought experiment. He imagines a situation, called the original position, in which it's *impossible* to choose in a directly self-serving way and then asks what people would choose. Choosing in a straightforwardly selfish manner is impossible in the original position, because one chooses from behind a *veil of ignorance*. That is, the parties in the original position must negotiate an agreement about how their society will be organized without knowing their own races, genders, ethnic backgrounds, social positions, economic statuses, or the nature and extent of their natural talents. The idea, then, is that the negotiators have been denied all of the information they could use to *bias* the agreement in their respective favors. The decision makers are expected to choose rationally and selfishly, but because they are choosing from behind a veil of ignorance, the kind of social structure they choose is, according to Rawls, necessarily fair and just. Agreeing on a social structure from behind a veil of ignorance is rather like using the "I cut, you choose" method for dividing a piece of cake. The fairness emanates from the decision procedure rather than from the goodwill of the decision makers.

This core idea (modeling social choice as bias-free selfish choice) was developed independently, and slightly earlier, by the Hungarian economist John Harsanyi (1953, 1955), who would later win the Nobel Prize in economics for his contributions to game theory. Harsanyi, unlike Rawls, saw his version of the original position as providing a rational grounding for utilitarianism. Harsanyi imagined people choosing their society's organizing principles while not knowing which positions in society they would occupy (rich or poor, etc.) but knowing that they would have an *equal probability* of occupying any position in society. Given this assumption, if people are utility maximizers (each seeking to maximize his/her own happiness), the decision makers will choose a society organized so as to maximize utility, one that is as happy as possible overall. (This maximizes both the *total* and the *average* amount of happiness, assuming that the population size is fixed.)

Rawls, however, argued for a very different conclusion about the kind of society that selfish people in the original position would choose. Rawls says that the original positioners would choose a society organized by a "maximin" principle rather than a utilitarian principle. The maximin principle ranks societies based solely on the status of the society's least well-off person. According to this principle, one's preferences for one societal arrangement over another will be based entirely on the "worst-case scenario" within each arrangement. Rawls acknowledges that this is not, in general, a good decision rule, as illustrated by the following example.

Suppose that you are buying a car, but in the following unusual way. You must buy a lottery ticket that will give you one car randomly chosen from a lot of a thousand cars. Ticket A takes you to a lot that has a thousand mediocre cars. On a scale of 1 to 10, each of these cars is a 4. If you buy ticket A, you get one of those. Ticket B takes you to a lot that also has a thousand cars. This lot has 999 cars that score a perfect 10

on your scale; however, it also has one car that scores a 3. So if you buy ticket B, you have a 99.9 percent chance of getting your dream car, but you have a 0.1 percent chance of getting a car that's okay but slightly worse than the one you're guaranteed to get with ticket A. Which do you choose? Obviously, you would choose ticket B. However, according to the "maximin" rule, you would choose ticket A, because the worst-case scenario in buying ticket A is better than the worst-case scenario in buying ticket B. Not so smart.

The problem with the maximin rule is that it's maximally *risk averse*. Rawls agrees that such risk aversion is not appropriate in general (e.g., when buying cars by lottery), but he argues that it *is* appropriate for people who are choosing their society's organizing principles without knowing which positions in society they will occupy. Rawls thinks that life in a utilitarian society might be "intolerable" (pp. 156, 175). If you're randomly plopped into a utilitarian society, Rawls warns, you could end up as a slave. Such outcomes are so bad that no one would choose to take that risk. Instead, people choosing from behind the veil of ignorance would use the maximin rule, favoring the society with the best worst-case scenario. Rawls makes this argument with respect to what he calls "basic liberties." Rather than leave the allocation of liberties up to utilitarian calculations, the people in the original position would choose principles that directly secured "basic liberties." He makes the same kind of argument about educational/economic opportunities and about economic outcomes. Here, too, says Rawls, the worst-case scenario in a utilitarian society could be so bad that it's not worth taking the risk.

First, let's note that Rawls's formal argument depends on the error described in the last chapter, confusing wealth and utility. More specifically, Rawls assumes that the people in the original position would make the same mistake that he makes. Once again, the reason for favoring the maximin rule is that the worst-case scenario in a utilitarian society might be "intolerable." It's not hard to see how the worst-case scenario in a *wealthitarian* society would be intolerable. Maximizing GDP might require some oppression, but, as explained earlier, it's simply not plausible that making the world as happy as possible could, in the real world, require oppression. Human psychology would have to be completely rewired such that the suffering caused by being a slave is smaller than the benefits one derives from owning a slave, and so on. (Once again, would you give up half your life to slavery in order to have a slave for the other half? Could you imagine a situation in the real world in which this is a tough decision?)

That's Rawls's first mistake. (I don't think this is a *rationalization*. I think it's just a mistake.) But now let's suppose that Rawls is right and that life in a maximally happy society could be, for some, "intolerable." Even if you make this implausible assumption, Rawls's argument still doesn't work. Once again, Rawls's maximin rule evaluates each societal arrangement based solely on its worst-case scenario—the quality of life of that society's least well-off person. In other words, Rawls assumes that people will be maximally risk averse so long as there are intolerable outcomes in the mix. But as Harsanyi (1953, 1955) and others have pointed out, that is simply not a reasonable assumption. Every time you get in a car, you increase your risk of being horribly maimed in a car accident, an outcome that most of us would regard as "intolerable" in Rawls's sense. And yet we accept such risks for things as trivial as a late-night pint of ice cream. (You might point out that one can be horribly maimed by staying home. For example, the roof could collapse. Thus, the worst-case scenario is actually the same whether you make your ice cream run or not. That's fine, but then you have to apply the same logic to Rawls's argument. Life could be "intolerable" even in a society governed by maximin. Your roof could collapse.)

To prime the risk aversion pump, Rawls adds an unnecessary twist to his version of the original position. In Harsanyi's version, if you recall, the decision makers choose while knowing that they will have an *equal probability* of occupying each position in society. Rawls, however, does something different. He assumes that people in the original position have no information at all about the range of possible outcomes and their associated odds, leaving them in a state of complete actuarial ignorance. In this state of maximal ignorance, Rawls argues, the people in the original position would, and should, be highly risk averse. ("Anything could happen!") In technical terms, Rawls makes the original position an *ambiguous* situation rather than merely an *uncertain* one.

Why does Rawls make the decision in the original position maximally ambiguous? Why not simply assume, as Harsanyi does, that the people in the original position know the range of possible social positions and know that they have equal odds of landing in any one of them? Rawls addresses this issue, and as far as I can tell, his argument is completely circular. He *defines* the original position as one in which people have no information about the attendant probabilities, and then argues, on that assumption, that they should not rely on probability estimates, because, really, they have no way of knowing what the probabilities are (pp. 155, 168–169). As Harsanyi (1975) points out, even under this assumption of complete ignorance, an assumption of equal probability for all outcomes would be far more rationally defensible than the assumption that the worst outcome has an effective probability of 100 percent—the assumption built into the maximin rule. But we can put that aside. Why, in the first place, should Rawls define the original position as one in which the outcome probabilities are unknown? The whole point of the original position is to constrain the choice so that

the decision makers are effectively impartial. To be impartial is to give equal weight to each person's interests. Thus, it makes perfect sense to *define* the original position as one in which each chooser knows that she has an equal probability of occupying each position in society. This would in no way bias people's choices. On the contrary, it embodies the ideal of impartiality in the clearest possible way.

As far as I can tell, Rawls makes the probabilistic structure of the original position maximally ambiguous for reasons that have nothing to do with justice or fairness or impartiality. As far as I can tell, this is just a fudge, an ad hoc attempt to make his intuitively correct answer more plausible. Coming into the world of political philosophy, Rawls has no particular reason to adopt an extreme theory of risk aversion. But once he's committed to the original position as a device for working out a theory of justice—which is a very nice idea—he suddenly finds himself in an awkward position. He wants a society in which priority is given, as a matter of first principles, to people with the worst outcomes. But this desire, filtered through the logic of his thought experiment, requires Rawls's hypothetical selfish decision makers to be inordinately preoccupied with the worst outcomes as they make their self-interested choices. That is, he needs them to be inordinately risk averse. And thus, to get his desired result, Rawls adds a layer of gratuitous ambiguity to his thought experiment to make extreme risk aversion seem more plausible.

As in Kant's case, this kind of finagling makes it clear what Rawls is really up to. He's not starting with first principles and then following them to their logical conclusions. He knows where he wants the argument to go, and he's doing everything he can to get it there.

Thus, Rawls's well-intentioned rationalizing illustrates two points. First, it's another nice example of what happens when very smart people are determined to vindicate their moral emotions through reasoning. Second, it suggests that Harsanyi may be right. If you run the original-position thought experiment properly by (a) not confusing wealth and utility, (b) not assuming that people are inordinately risk averse, and (c) not making the hypothetical decision gratuitously ambiguous, you just might end up with a utilitarian conclusion. In other words, if you replace the happiness assumption with an assumption favoring choice, you end up with utilitarianism, because impartial people, with no ideological commitments, will naturally choose a society that maximizes their prospects for happiness.

334 **wasn't always a liberal:** In my youth I was something of a libertarian conservative. My libertarian claim to fame: In my senior year of high school I won third prize in an Ayn Rand essay contest. However, by the time the prize came through, I was already starting to change my mind. I shared my doubts with the woman who called to congratulate me. This did not go over well.

334 **Jonathan Haidt:** Haidt (2001, 2007, 2012).

336 **"Emotional Dog":** Haidt (2001).

336 **this characterization of his view:** According to Haidt, moral reasoning plays an important role in his landmark theory of moral psychology, the Social Intuitionist Model (SIM) (Haidt, 2001). Whether this is true depends on what counts as moral "reasoning" (Paxton and Greene, 2010).

According to the SIM, moral judgment works as follows: Moral judgments are, in general, caused by moral intuitions, and when we engage in moral reasoning, our reasoning is typically deployed post hoc to justify the moral judgments that we have already made on an intuitive basis. (See discussion of moral rationalization earlier in this chapter.) Haidt says that people do sometimes engage in private moral reasoning, but that this is ''rare, occurring primarily in cases in which the intuition is weak and processing capacity is high'' (p. 819). This is why I say that, according to Haidt, moral reasoning plays a minor role in moral life.

However, there are two additional psychological processes to consider, the ones that put the "Social" in the SIM. First, according to the SIM, Person A's overtly making a moral judgment can influence Person B's moral intuitions, which can in turn influence Person B's moral judgment. Haidt calls this "social persuasion." This is clearly not moral reasoning, as there is no argument, just an intuitive response to observing another's judgment or behavior. Second—and this is the key part—Haidt says that people engage in "reasoned persuasion." Here, Person A provides a verbal justification for her judgment; Person B hears this justification; and this modifies his moral intuitions, which in turn influences his moral judgment. Haidt calls this "reasoned persuasion," but that label, I think, is misleading. Here, Person A influences Person B's judgment by modifying Person B's *feelings* (automatic settings) and not by engaging Person B's capacity for explicit reasoning (manual mode). Here the "reasons" that Person A produces function like a song that succeeds in moving Person B.

Haidt believes that this process is widespread and highly influential (which it may be). This is why he says that moral reasoning plays an important role in moral life. But, as I've said, I don't think that this qualifies as "moral reasoning." And that is why I say, over Haidt's protests, that his view is not one according to which moral reasoning plays a major role. According to the SIM, I cannot change your mind on a moral issue (such as gay marriage, abortion, or eating animals) without first changing your feelings. I cannot appeal directly to your capacity for reasoning and thus cause you to override your feelings. I think that this picture of moral psychology is incorrect. This point is illustrated by an experiment in which my collaborators and I

used a rather abstract argument to persuade people (at least temporarily) to accept a counterintuitive moral conclusion (Paxton, Ungar, and Greene, 2011).

336 **Some liberals say . . . some social conservatives believe:** Here are two examples: http://www.libchrist.com/other/abortion/choice.html; http://k2globalcommunicationsllc.wordpress.com/2012/08/28/abortion-nihilist-argument-eliminate-poverty-kill-the-poor. See also John Paul II (1995).

336 **liberals have impoverished moral sensibilities:** Haidt and Graham (2007); Graham, Haidt, et al. (2009); Haidt (2012).

337 **doubts about this six-part theory . . . important aspect . . . well-supported:** The survey data (Graham et al., 2009, 2011) that Haidt uses to support his theory (the original version with five foundations) show an enormous division between two clusters: the care-fairness cluster and the loyalty-authority-sanctity cluster. There is, by contrast, relatively little evidence for a two-way division within the first cluster or a three-way division within the second cluster, and what evidence there is can be accounted for by the fact that the surveys used to collect these data were designed with five clusters in mind. To provide strong evidence for a five-factor (or six- or n-factor) theory of morality, one would have to use a "bottom-up" approach, testing the theory using testing materials that were not designed with any particular theory in mind. Haidt says that, to a first approximation, the moral world has five (or six) "continents." In Haidt's data, I see evidence for two continents, which may or may not have two or three interesting bulges.

337 **questions like these:** Graham, Haidt, et al. (2009) posed these questions in a different way ("How much money would it take for you to . . . ?").

338 **Bentham and Kant:** I think that Haidt's (2012) psychological portrait of Kant (p. 120) is off the mark. Kant might have had some autistic tendencies, and he was certainly a "systematizer," but he was very authoritarian and no stranger to moral disgust. He was also, not incidentally, very religious.

338 **WEIRD:** See Henrich, Heine, et al. (2010).

338 **predicting what liberals will say . . . vice versa:** Graham, Nosek, et al. (2012).

339 **dismisses . . . science:** Mooney (2012).

339 **"real" Americans:** Devos and Banaji (2005).

339 **45 percent of Republicans believe:** Condon (April 21, 2011).

339 **little respect for . . . United Nations:** Wike (September 21, 2009).

339 **Muslim American . . . should not be trusted:** Arab American Institute (August 22, 2012).

340 **If Iranians . . . want to protest:** Swami (June 15, 2009).

340 **minority . . . report believing in God:** European Commission (2005).

340 **lowest crime rates . . . happiness:** Economic Intelligence Unit (2005); United Nations Office of Drugs and Crime (2011); United Nations (2011); Ingelhart, Foa, et al. (2008). Murder rates, educational attainment, and test scores: World Values Survey on happiness.

340 **"yang" . . . "yin":** Haidt (2012), 294.

340 **no Republicans hold elected office:** Based on my own online searches and confirmed by Henry Irving, Republican city committee chair, Cambridge, MA (personal communication, March 24, 2013).

340 **bonds are rated AAA:** http://www.cambridgema.gov/citynewsandpublications/news/2012/02/cambridgemaintainsraredistinctionofearningthreetriplearatings.aspx.

340 **social conservatives are very good at:** Putnam (2000); Putnam and Campbell (2010).

342 **"Spread my work ethic":** Haidt (2012), 137.

342 **Refusing to buy . . . more harm than good:** Nicholas D. Kristof, "Where Sweatshops Are a Dream." *New York Times*, January, 14, 2009.

343 **"welfare queens":** The term was coined by Ronald Reagan during his 1976 presidential campaign: "'Welfare Queen' Becomes Issue in Reagan Campaign." *New York Times*, February 15, 1976.

343 **former slave states:** Lind (2012).

343 **government hands off my Medicare:** Krugman (July 28, 2009).

343 **billionaires . . . lower rate than their secretaries:** Tienabeso (January 25, 2012); Buffett (August 14, 2011).

344 **reliable returns:** Yes, such donors might benefit enormously from a Romney administration's policies, but the odds that any single donor's donation would sway the election are extremely small.

344 **endorses . . . utilitarianism:** Haidt (2012) endorses, for the purposes of making policy, what he calls "Durkheimian utilitarianism" (p. 272), which is utilitarianism that accounts for the value of conservative social institutions such as religion. Durkheimian utilitarianism is actually just utilitarianism wisely applied. Nevertheless, the point is worth making because not all self-styled utilitarians appreciate the value of conservative social institutions. Mill (1885), however, certainly did, as explained, for example, in his essay "The Utility of Religion."

345 **"I don't know":** Haidt (2012), 272.

345 **"90 percent chimp and 10 percent bee":** Ibid., xv.

346 **moral reasoning . . . ineffective, though not completely:** See Paxton, Ungar, and Greene (2011).

346 **a good argument can change the shape of things:** An alternative analogy: A good argument is like a piece of technology. Few of us will ever invent a new piece of technology, and on any given day it's unlikely that we'll adopt one. Nevertheless, the world we inhabit is defined by technological change. Likewise, I believe that the world we inhabit is a product of good moral arguments. It's hard to catch someone in the midst of reasoned moral persuasion, and harder still to observe the genesis of a good argument. But I believe that without our capacity for moral reasoning, the world would be a very different place. See also Pinker (2011), chaps. 9–10; Pizarro and Bloom (2003); Finnemore and Sikkink (1998).

346 **"I have been tormenting":** Bentham (1978).

346 **"But it would be a mistake":** Mill (1895), 1–2.

# 第12章 傻瓜道德模式之外：6条规则

349 **Chekhov:** See introduction, "man will become."

350 **consult, but do not trust, your moral instincts:** This rule is available in bumper sticker form: "Don't believe everything you think." Available at www.northernsun.com.

350 **manual mode . . . Me ahead of Us:** Valdesolo and DeSteno (2007).

352 **more important than saving someone's life:** Likewise, few of us can honestly say that animals should suffer enormous pain because pork is tastier than tofu or because an extra dollar is too much to spend on a cruelty-free cheeseburger (not widely available, but only for lack of demand).